

EXHIBIT 2

Part 1

Filed Under Seal

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

Daily News, LP; Chicago Tribune Company, LLC;
Orlando Sentinel Communications Company, LLC;
Sun-Sentinel Company, LLC; San Jose Mercury-
News, LLC; DP Media Network, LLC; ORB
Publishing, LLC; and Northwest Publications, LLC,

Plaintiffs,

v.

MICROSOFT CORPORATION, OPENAI, INC.,
OPENAI LP, OPENAI GP, LLC, OPENAI, LLC,
OPENAI OPKO, LLC, OPENAI GLOBAL, LLC,
OAI CORPORATION, LLC, and OPENAI
HOLDINGS, LLC,

Defendants.

Civil Action No. 1:24-cv-
3285

FIRST AMENDED COMPLAINT

JURY TRIAL DEMANDED

Plaintiffs Daily News, LP (the “New York Daily News”); Chicago Tribune Company, LLC, (the “Chicago Tribune”); Orlando Sentinel Communications Company, LLC (the “Orlando Sentinel”); Sun-Sentinel Company, LLC (the “Sun-Sentinel”); San Jose Mercury-News, LLC (the “Mercury News”); DP Media Network, LLC (the “Denver Post”); ORB Publishing, LLC (the “Orange County Register”); and Northwest Publications, LLC (the “Pioneer Press”) (collectively the “Publishers”), by their attorneys Rothwell, Figg, Ernst & Manbeck, P.C., for their ~~complaint~~ [First 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”), allege as follows:

I. NATURE OF THE ACTION

1. This lawsuit arises from Defendants purloining millions of the Publishers’ copyrighted articles without permission and without payment to fuel the commercialization of their

generative artificial intelligence (“GenAI”) products, including ChatGPT and Copilot. Although OpenAI purported at one time to be a non-profit organization, its recent \$90 billion valuation underscores how that is no longer the case. ChatGPT, along with Microsoft Copilot (formerly known as Bing Chat) has also added hundreds of billions of dollars to Microsoft’s market value. Defendants have created those GenAI products in violation of the law by using important journalism created by the Publishers’ newspapers without any compensation.

2. To build and operate GenAI products, Defendants use computers. They pay for those computers. Defendants use specialized chips. They pay for those chips. Defendants use electricity to operate the computers. They pay for that electricity. Defendants employ programmers and other technical employees. They pay those programmers and other employees. They build and operate their GenAI products in expensive and highly sophisticated facilities. They pay for those facilities.

3. Defendants also need high quality content in order to make their GenAI products successful. OpenAI’s founder, Sam Altman, conceded in testimony to the House of Lords in the U.K. that his company could not make commercially viable GAI products without using copyrighted material:

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

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

4. Despite admitting that they need copyrighted content to produce a commercially viable GenAI product, Defendants contend that they can fuel the creation and operation of these products with the Publishers' content without permission and without paying for the privilege. They are wrong on both counts, as this lawsuit will prove.

5. In recent years, the biggest threat to local news generally, and local newspapers in particular, has been the development of the internet and the theft of newspapers' content and the consequent siphoning of advertising revenue. The newspapers that survived this historic transformation in news delivery were able to do so in part because they continued to provide content that their readers found to be informative, entertaining, and valuable.

6. Microsoft and OpenAI simply take the work product of reporters, journalists, editorial writers, editors and others who contribute to the work of local newspapers—all without any regard for the efforts, much less the legal rights, of those who create and publish the news on which local communities rely. The Publishers are regional and local news organizations that provide reporting critical for the neighborhoods and communities that form the very foundation of our great nation. The Publishers' journalism ensures that government and power are accountable to the people, that taxpayers get what they pay for, that there is justice for all, and that citizens receive timely information critical for their daily lives. The Publishers also connect people and build consensus by keeping their readers informed.

7. The Publishers have spent billions of dollars sending real people to real places to report on real events in the real world and distribute that reporting in their print newspapers and on their digital platforms. Yet Defendants are taking the Publishers' work with impunity and are using the Publishers' journalism to create GenAI products that undermine the Publishers' core businesses by retransmitting "their content"—in some cases verbatim from the Publishers'

paywalled websites—to their readers. As if plagiarizing the Publishers’ work were not enough, Defendants’ products are often subject to “hallucinations” where those products malign the Publishers’ credibility by falsely attributing inaccurate reporting to the Publishers’ newspapers. Beyond just profiting from the theft of the Publishers’ content, Defendants are actively tarnishing the newspapers’ reputations and spreading dangerous disinformation.

8. This issue is not just a business problem for a handful of newspapers or the newspaper industry at large. It is a critical issue for civic life in America. Indeed, local news is the bedrock of democracy and its continued existence is put at risk by Defendants’ actions.

9. The late Speaker of the House, Thomas P. “Tip” O’Neill, famously remarked that all politics is local. That is as true today as when he said it. Presidential primaries begin in Iowa, New Hampshire, and South Carolina, and their outcomes are often decided in states such as Florida—where two of the Publishers operate. Congressmen and Senators typically begin their political careers in local city councils, mayoral offices, and state legislatures. When major national news outlets, like CNN, MSNBC, and Fox News, need insights into local politics, they frequently rely on local news reports from the Publishers and others, and often turn to reporters from local newspapers for on-air commentary. Politics is local, and local news is an integral part of our national conversation.

10. Of course, politics is a part of life, but not all of it. When residents of a city or town want to know what is happening in their community, they look to local news outlets for information on arts and entertainment, sports, weather, and civic events. Who won the local high school or college game? What’s playing at local theaters? Will the weather be nice this weekend? What will the latest city council resolutions on taxes, education, and public transportation mean for my neighborhood? Local news—and specifically local newspapers—provide in one place the

information that scores of millions of Americans rely on for the trustworthy and up-to-date information they need to organize and plan their lives. And local news provides much more, including job availabilities, house and apartment listings, obituaries, and coverage of other topics that national news, and even some of the largest city papers, choose not to cover.

11. The impact of local newspapers—or, unfortunately for many Americans, the absence of them—has important social consequences. Northwestern University’s Local News Initiative explained the impact of the newspaper divide on American life²:

Economically struggling and traditionally underserved communities—where residents need journalists providing transparency and oversight of local government and business decisions—are the ones most likely to lose a news organization and be overlooked by funders looking to invest in both for-profit and nonprofit news operations. That loss of local journalism exacerbates political, cultural and economic divisions between and within communities.

12. The Publishers are local newspapers serving many of the largest metropolitan areas of the country (including New York, Chicago, Orlando, and San Jose) as well as smaller cities, towns and regions (such as Hartford, Connecticut; Allentown, Pennsylvania; and Norfolk, Virginia). Collectively, the Publishers are working to narrow the news divide, and to fill in the “news deserts” that have arisen in many states and regions.

13. These surviving newspapers—increasingly a rare breed in America—now face a new threat: the GenAI products offered by Defendants Microsoft and OpenAI. In their GenAI products’ formative stages, Microsoft and OpenAI “scrape” the newspapers’ content, copying it onto their systems. The scraped content becomes part of the raw material on which Microsoft and

² P.M. Abernathy, *The State of Local News: The 2022 Report, Executive Summary*, NORTHWESTERN UNIVERSITY (June 29, 2022), <https://localnewsinitiative.northwestern.edu/research/state-of-local-news/2022/report/>.

OpenAI train their GenAI systems. That is, they copy the informative, entertaining, and valuable content and style that newspapers (including the Publishers' newspapers) offer their own paying subscribers. At the end of the process, the Microsoft and OpenAI GenAI systems offer their users content that is identical to, or a slightly masked version of, the newspapers' content.

14. Microsoft and OpenAI bizarrely claim they are entitled to copy and use any written product on which they want to train their GenAI systems. Microsoft and OpenAI also say even after their systems are trained, they are entitled to copy local newspapers day in and day out, and store the newspapers' content on their servers, as source materials for their GenAI products' output. They even go so far as to say that their GenAI products are allowed to copy and provide back to their users verbatim reprints of local news articles, editorials, reviews, and stories that the local reporters and editors spend their time, resources, and in some cases even their physical safety, to write and publish. Microsoft and OpenAI have gone so far as to say that, if anyone should be held liable for the improper use of local newspaper content, it should be the unsuspecting users of Microsoft's and OpenAI's own products.

15. Since the founding of the United States, the Constitution has granted Congress the authority to "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." Since our Nation's founding in 1789, Congress has seen fit to protect a newspaper's written material, to prohibit the copying of that material without the newspaper's permission, and to require that anyone who uses that material provide compensation to the newspaper. Microsoft and OpenAI are ardent defenders of these principles when it comes to their own products, using paywalls and "shrinkwrap" licenses to protect their GenAI systems and the other products they offer. We can be quite sure that if newspapers throughout the country decided to use pirated copies of Microsoft

Windows or Word, or OpenAI's Chat GPT-4, Microsoft and OpenAI would be quick to assert their rights. Yet when the shoe is on the other foot, Microsoft and OpenAI say that newspaper content is not entitled to the same protection.

16. Plaintiffs seek no more than what Microsoft and OpenAI claim for themselves: recognition that newspapers have legal rights in their content, that Microsoft and OpenAI are legally required to respect those rights, and that Microsoft and OpenAI owe the newspapers compensation for their unlawful use of protected newspaper content to date. Both Microsoft and OpenAI are well aware of the value of the newspaper content they have exploited. Microsoft's deployment of Copilot (formerly known as Bing Chat), its GenAI tool that uses the Bing search index, has helped boost its market capitalization by a trillion dollars in the past year alone. OpenAI, which used to pretend it was a non-profit organization, now has a market capitalization of over \$90 billion. Yet they both continue to deny that they owe anything to even a single newspaper whose content they have copied, and continue to copy, to build and operate their GenAI products.

17. This lawsuit is not a battle between new technology and old technology. It is not a battle between a thriving industry and an industry in transition. It is most surely not a battle to resolve the phalanx of social, political, moral, and economic issues that GenAI raises. This lawsuit is about how Microsoft and OpenAI are not entitled to use copyrighted newspaper content to build their new trillion-dollar enterprises, without paying for that content. As this lawsuit will demonstrate, Defendants must both obtain the Publishers' consent to use their content and pay fair value for such use.

II. JURISDICTION AND VENUE

18. 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.

19. Jurisdiction over Microsoft and OpenAI is proper because they have purposely availed themselves of the privilege of conducting business in New York. A substantial portion of Microsoft and OpenAI's widespread infringement and other unlawful conduct alleged herein occurred in New York, including the distribution and sales of Microsoft and OpenAI's Generative Pre-training Transformer ("GPT")-based products like ChatGPT, ChatGPT Enterprise, Copilot, Azure OpenAI Service, Microsoft 365 Copilot, 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.

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

21. Plaintiff Daily News, LP (the “New York Daily News”) is a Delaware limited partnership with offices located in New York. The New York Daily News publishes digital and print products, including its core news product, The New York *Daily News*, which is available on its mobile application, on its website (www.nydailynews.com), and as a printed newspaper. The New York Daily News owns over 9,000 registered copyrights for its newspaper issues, including those set forth in Exhibit A (“New York Daily News Works”).

22. Plaintiff Chicago Tribune Company, LLC (the “Chicago Tribune”) is a Delaware limited liability company with a principal place of business located in Illinois. The Chicago Tribune publishes digital and print products, including its core news product, The *Chicago Tribune*, which is available on its mobile application, on its website (www.chicagotribune.com), and as a printed newspaper. The Chicago Tribune owns over 16,000 registered copyrights for its newspaper issues, including those set forth in Exhibit B (“Chicago Tribune Works”).

23. Plaintiff Orlando Sentinel Communications Company, LLC (the “Orlando Sentinel”) is a Delaware limited liability company with a principal place of business located in Florida. The Orlando Sentinel publishes digital and print products, including its core news product, The *Orlando Sentinel*, which is available on its mobile application, on its website (www.orlandosentinel.com), and as a printed newspaper. The Orlando Sentinel owns over 7,000 registered copyrights for its newspaper issues, including those set forth in Exhibit C (“Orlando Sentinel Works”).

24. Plaintiff Sun-Sentinel Company, LLC (the “Sun-Sentinel”) is a Delaware limited liability company with a principal place of business located in Florida. The Sun-Sentinel publishes digital and print products, including its core news product, The *Sun-Sentinel*, which is available on its mobile application, on its website (www.sun-sentinel.com), and as a printed newspaper. The

Sun-Sentinel owns over 8,000 registered copyrights for its newspaper issues, including those set forth in Exhibit D (“Sun-Sentinel Works”).

25. Plaintiff San Jose Mercury-News, LLC (the “Mercury News”) is a California limited liability company with a principal place of business located in California. The Mercury News publishes digital and print products, including its core news product, The *Mercury News*, which is available on its mobile application, on its website (www.mercurynews.com), and as a printed newspaper. The Mercury News owns over 10,000 registered copyrights for its newspaper issues, including those set forth in Exhibit E (“Mercury News Works”).

26. Plaintiff DP Media Network, LLC (the “Denver Post”) is a Delaware limited liability company with a principal place of business located in Colorado. The Denver Post publishes digital and print products, including its core news product, The *Denver Post*, which is available on its mobile application, on its website (www.denverpost.com), and as a printed newspaper. The Denver Post owns over 5,000 registered copyrights for its newspaper issues, including those set forth in Exhibit F (“Denver Post Works”).

27. Plaintiff ORB Publishing, LLC (the “Orange County Register”) is a Delaware limited liability company with a principal place of business located in California. The Orange County Register publishes digital and print products, including its core news product, The *Orange County Register*, which is available on its mobile application, on its website (www.oregister.com), and as a printed newspaper. The Orange County Register owns over 9,000 registered copyrights for its newspaper issues, including those set forth in Exhibit G (“Orange County Register Works”).

28. Plaintiff Northwest Publications, LLC (the “Pioneer Press”) is a Delaware limited liability company with a principal place of business located in Minnesota. The Pioneer Press publishes digital and print products, including its core news product, The *Pioneer Press*, which is

available on its mobile application, on its website (www.twincities.com), and as a printed newspaper. The Pioneer Press owns over 9,000 registered copyrights for its newspaper issues, including those set forth in Exhibit H (“Pioneer Press Works”).

29. 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.

30. Microsoft has described its relationship with the OpenAI Defendants as a “partnership.” This partnership has included contributing and operating the cloud computing services used to copy the New York Daily News Works, the Chicago Tribune Works, the Denver Post Works, the Orlando Sentinel Works, the Sun-Sentinel Works, the Mercury News Works, the Orange County Register Works, and the Pioneer Press Works (collectively the “Publishers’ Works”) and train the OpenAI Defendants’ GenAI models. It has also included, on information and belief, substantial technical collaboration on the creation of those models. Microsoft possesses copies of, or obtains preferential access to, the OpenAI Defendants’ latest GenAI models that have been trained on and embody unauthorized copies of the Publishers’ Works. Microsoft uses these models to provide infringing content and, at times, misinformation to users of its products and online services. During a quarterly earnings call in October 2023, Microsoft noted that “more than 18,000 organizations now use Azure OpenAI Service, including new-to-Azure customers.”

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

32. 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.

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

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

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

36. Defendant OpenAI OpCo, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street, San Francisco, California. OpenAI OpCo,

LLC is a wholly owned subsidiary of OpenAI, Inc. and has facilitated and directed OpenAI's mass infringement and unlawful exploitation of the Publishers' Works through its management and direction of OpenAI, LLC.

37. 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.

38. 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.

39. 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 Publishers**

40. Plaintiff Daily News, LP is the publisher of the New York *Daily News*, which was the first newspaper in America to publish in the now-familiar tabloid format. Established in 1919, and originally called the *Illustrated Daily News*, the *Daily News* has long been renowned especially for its award-winning photojournalism, its iconic headlines and front-page visuals, and its in-depth

coverage of local events, politics, and sports. The *Daily News* received its most recent Pulitzer Prize for Public Service in 2017 for (in the words of the Pulitzer organization) its coverage of “widespread abuse of eviction rules by the police to oust hundreds of people, most of them poor minorities.” In 2007, the *Daily News*’ editorial board won the Pulitzer Prize for a series of editorials highlighting the declining health of the thousands 9/11 first responders. Following the *Daily News*’ reporting, Congress passed the James Zadroga 9/11 Health and Compensation Act, Pub. L. 111-347, which established a federal program to monitor the health of 9/11 first responders, and to provide aid to those responders in need.

41. Plaintiff Chicago Tribune Company, LLC is the publisher of the *Chicago Tribune*. Established in 1847, eight years later the *Tribune* emerged as one of the first and leading voices for “Free Soil” and the abolition of slavery in the United States. The *Tribune* was an early promoter of Abraham Lincoln’s 1860 presidential candidacy, and ardently supported his administration. Following the Civil War, the *Tribune* remained a leading voice in the industrial Midwest during the late 19th and 20th centuries. Its impact on national and international news has been significant. In 1919, the *Tribune* was the first to obtain and publish the text of the Treaty of Versailles, which ended the first world war. On May 1, 1974, the *Tribune* was the first newspaper to publish a full transcript of President Richard Nixon’s Oval Office tapes, leading to his resignation in August of that year. In 1933, the *Tribune*’s sports editor created the first Major League Baseball all-star game. The *Tribune* staff and reporters have won 28 Pulitzer Prizes, including most recently in 2022 (in the Local Reporting category, for a series on Chicago’s deficient building- and fire-safety code enforcement); in 2017 (for Feature Photography, depicting a boy and his mother as their lives moved forward after he survived a shooting); and in 2017 (Public Service, for reporting on harmful

pharmacy practices, and particularly noting that the reporting led to reforms that prevented future harm).

42. Plaintiff San Jose Mercury-News, LLC is the publisher of the *San Jose Mercury News*. The *Mercury News* traces its roots to the California Gold Rush and has been covering the region now known as Silicon Valley under various names since 1851. It is one of the oldest continuously published newspapers in California. It has also been the local news publisher of record for Silicon Valley, the heart of U.S. tech innovation. For more than half a century, the *Mercury News* has been a leading source of news and information regarding the growth and development of the high-technology industry and its impact on the U.S. economy. The *Mercury News* has received two Pulitzer Prizes: one in 1986 for international reporting for a series of stories that documented massive transfers of wealth abroad by President Ferdinand Marcos and his associates and played a role in the Philippine president's downfall; and one in 1990 for coverage of the San Francisco earthquake. Consistent with its focus on technology, the *Mercury News* was among the first daily newspapers with an online presence, the first daily newspaper to put its entire content on its website, and the first daily newspaper to use its website to announce breaking news.

43. Plaintiff Orlando Sentinel Communications Company, LLC is the publisher of the *Orlando Sentinel*. The *Sentinel* is the successor to the *Orange County Reporter*, which issued its first edition in 1876. Currently, the *Orlando Sentinel* covers a diverse part of Florida that includes tourism, construction and development, agriculture, Disney World, and NASA's Kennedy Space Center. From 2001 through 2022, the Orlando Sentinel published a weekly Spanish-language paper, *El Sentinel*. In addition to chronicling the rapid transformation of central Florida from a rural agricultural region to a global vacation and space hub, the *Orlando Sentinel* has received accolades for its reporting, having won three Pulitzer Prizes. The prizes ranged from investigative

reporting for chronicling racial profiling and unjust property seizures by a local sheriff's office (1993) and two Pulitzers for editorial writing regarding lax growth management (1982) and predatory lending practices in Florida (2000). More recently, the *Orlando Sentinel* was named a Pulitzer finalist in 2017 for its breaking news coverage of the Pulse nightclub massacre. It also was named a Pulitzer finalist in 2013 in local reporting for a hazing tragedy at Florida A&M University.

44. Plaintiff Sun-Sentinel Company, LLC is the publisher of the *South Florida Sun Sentinel*, a sister newspaper of the *Orlando Sentinel* that serves southeast Florida. As with the *Orlando Sentinel*, the *Sun Sentinel* has been a newspaper of record for one of the most dynamic, growing, and diverse regions in the country, Broward and Palm Beach counties, serving a population of 3.5 million. In 2013, the *Sun Sentinel* won its first Pulitzer Prize, awarded in the category of Public Service journalism, for its series regarding reckless conduct by off-duty police. It won its second Pulitzer Prize in the same category six years later for its coverage of the mass shooting at Marjory Stoneman Douglas High School. Its 10-month investigation of the Broward Sheriff's Office and Broward School District's actions before, during and after the massacre led to changes in state law regarding school safety and led to mass resignations and even prosecution of deputies at the tragic scene. A 2022 investigative series on sex trafficking in South Florida also led to bipartisan state legislation cracking down on sex traffickers and the hotels that facilitate it. The newspaper was a Pulitzer Prize finalist in 2006 for its reporting on federal government mismanagement during and following Hurricane Wilma. Its photojournalists and visuals staff have won international recognition for their photography and interactive graphics, and the features writing department has placed in national contests several times for chronicling South Florida life.

45. Plaintiff DP Media Network, LLC is the publisher of the *Denver Post*. Called the *Evening Post* when it was first published in 1892, the newspaper has been a leading news outlet in the Rocky Mountain region for over 125 years. Since 2000, the *Post* has received five Pulitzer prizes in a range of categories, and for a variety of subjects—a record that underscores the *Post*'s significance in a large and diverse region with few major metropolitan areas. The *Post* received Pulitzer Prizes in 2000 for Breaking News Reporting (the *Post*'s coverage of the Columbine High School Massacre); in 2010 for Feature Photography; in 2011 for Editorial Cartooning; in 2012 for Feature Photography; and in 2013 for Breaking News Reporting (the Aurora, Colorado shooting). During the same period, the *Post* was a Pulitzer Prize finalist twice, in 2007 (Breaking News Reporting for coverage of the region's back-to-back blizzards), and in 2015 (Explanatory Reporting for coverage of Colorado's marijuana laws).

46. Plaintiff ORB Publishing, LLC is the publisher of the *Orange County Register*, which serves the southern region of California centered on Orange and Los Angeles Counties. The *Register* has been known throughout its 120-year history for its iconoclastic stances on major issues, many of which foreshadowed dramatic changes in American political and social attitudes. In 1942, at the height of World War II, the *Register* published an editorial opposing the internment of Japanese citizens on legal and humanitarian grounds. The *Register* was one of a few newspapers to oppose the Iraq War from the outset, and it was an early opponent of efforts in California to prohibit same-sex marriage. The *Register* has received three Pulitzer Prizes: in 1985 for photography at the 1984 Olympics in Los Angeles; in 1989 for specialized reporting for coverage of the military in Southern California; and in 1996 for investigative reporting that uncovered fraud and unethical activity at a leading research university hospital and prompted key regulatory reforms. Since 2000, the *Register* has been a finalist for two Pulitzer Prizes: in 2004 recognizing

three *Register* reporters for their investigation of care at 26 local hospitals, and for the reporters' development of a "report card" to help patients make informed decisions; and in 2005 for reporting on lead-containing candy from Mexico and its effects on children. The *Register* continues to do important public-service journalism, including continued coverage of systemic weakness and fraud in rehab facilities throughout Southern California and ongoing reporting about misconduct and abuse in Olympic and college sports programs.

47. Plaintiff Northwest Publications, LLC is the publisher of the *St. Paul Pioneer Press*. The *Pioneer Press* is Minnesota's first newspaper, celebrating its 175th anniversary this year. Its earliest ancestor was the *Minnesota Pioneer*, first published in 1849, just two years after Minnesota was established as a territory and nine years before it became a state. Winner of three Pulitzer Prizes, the *Pioneer Press* has been known for investigative and local enterprise work, for its commitment to distinct local news and to St. Paul and area communities. Based in Minnesota's capital, in a highly competitive news market, the *Pioneer Press* reports on the Twin Cities of St. Paul and Minneapolis as well as parts of western Wisconsin.

48. The Publishers expend significant time and effort investigating and reporting local stories and rely on ad and subscription revenue to further this enterprise. Defendants' actions threaten the Publishers' continued efforts to provide American communities with quality, in-depth local journalism by copying, using, and adapting the Publishers' Works in connection with GenAI products without compensation to the Publishers. These actions deprive the Publishers of visits to their sites, decrease subscription revenue, and deprive the Publishers of licensing revenue.

49. To preserve the vitality of local journalism, the Publishers go to great lengths to protect their content. They routinely register copyrights in their content and provide copyright notices in connection with their works. Beyond the protections provided by the exclusive rights of

reproduction, adaptation, publication, performance, and display under copyright law, the Publishers use paywalls to protect some of their content and implement terms of service and terms of use that restrict the use of the content provided on their websites.³

50. The Publishers require that any third party that wishes to use their content obtain a license to do so. These licensing agreements allow the Publishers to control how third parties receive and display their content. The Publishers license their content only under narrowly tailored terms that provide explicit guardrails regarding how and to what extent third parties can use the licensed content.

51. Even under these licensing agreements, third parties are not permitted to “scrape” the content from the Publishers’ websites, as Defendants have done without permission. Instead, the Publishers retain control over how third parties access licensed content by requiring them to use specific channels to obtain the Publishers’ Works.

B. Defendants’ GenAI Products

1. A Joint Enterprise Based on Mass Copyright Infringement

52. 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.

³ See Central Terms of Service, TRIBUNE PUBLISHING, <https://www.tribpub.com/central-terms-of-service/> (last visited Apr. 17, 2024); Terms of Use, MEDIANEWS GROUP, <https://www.medianewsgroup.com/terms-of-use/> (last visited Apr. 17, 2024).

53. 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."⁴ 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."⁵

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

⁴ Greg Brockman & Ilya Sutskever, *Introducing OpenAI*, OPENAI (Dec. 11, 2015), <https://openai.com/blog/introducing-openai>.

⁵ *Id.*

55. 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.

56. 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.”⁶

57. 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.”⁷ But its effect was clearly to conceal the identity of the data OpenAI copied to train its latest models from rightsholders like the Publishers.

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

⁶ OPENAI, GPT-4 TECHNICAL REPORT (2023), <https://cdn.openai.com/papers/gpt-4.pdf>.

⁷ James Vincent, *OpenAI Co-Founder on Company’s Past Approach to Openly Sharing Research: ‘We Were Wrong’*, THE VERGE (Mar. 15, 2023), <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closedresearch-ilya-sutskever-interview>.

59. 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.

60. These commercial offerings have been immensely valuable for OpenAI. Over 80% of Fortune 500 companies are using ChatGPT.⁸ According to recent reports, in December 2023 OpenAI achieved \$2 billion in revenue and expects to double this figure to \$4 billion in revenue in 2025.⁹

61. 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 Publishers in the course of “training” the LLM.

62. Upon information and belief, all of the OpenAI Defendants have been either directly involved in or have directed, controlled, and profited from OpenAI’s widespread infringement and commercial exploitation of the Publishers’ Works. OpenAI, Inc., alongside Microsoft, controlled and directed the widespread reproduction, distribution, and commercial use of the Publishers’ Works perpetrated by OpenAI LP and OpenAI Global, LLC, through a series of

⁸ OpenAI, *Introducing ChatGPT Enterprise*, OPENAI (Aug. 28, 2023), <https://openai.com/blog/introducing-chatgpt-enterprise>.

⁹ Reuters, *OpenAI hits \$2 bln revenue milestone – FT*, REUTERS (Feb. 9, 2024), <https://www.reuters.com/technology/openai-hits-2-bln-revenue-milestone-ft-2024-02-09/>.

holding and shell companies that include OpenAI Holdings, LLC, OpenAI GP, LLC, and OAI Corporation, LLC., OpenAI LP and OpenAI Global, LLC were directly involved in the design, development, and commercialization of OpenAI's GPT-based products, and directly engaged in the widespread reproduction, distribution, and commercial use of the Publishers' Works. OpenAI LP and OpenAI Global, LLC also controlled and directed OpenAI, LLC and OpenAI OpCo, LLC, which were involved in distributing, selling, and licensing OpenAI's GPT-based products, and thus monetized the reproduction, distribution, and commercial use of the Publishers' Works.

63. 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 twofour ways.

64. 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 Publishers' intellectual property for the purpose of creating the GPT models that exploit and, in many cases, retain large portions of the copyrightable expression contained in those works.

65. 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.”

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

But beneath what OpenAI is putting out as large models, remember, the heavy lifting was done by the [Microsoft] Azure team to build the computer infrastructure. Because these workloads are so different than anything that's come before. So we needed to completely rethink even the datacenter up to the infrastructure that first gave us even a shot to build the models. And now we're translating the models into products.¹⁰

67. Microsoft built this supercomputer “in collaboration with and exclusively for OpenAI,” and “designed [it] specifically to train that company's AI models.”¹¹ 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.”¹² This system ranked in the top five most powerful publicly known supercomputing systems in the world.

68. 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

¹⁰ *First on CNBC: CNBC Transcript: Microsoft CEO Satya Nadella Speaks with CNBC's Jon Fortt on “Power Lunch” Today*, CNBC (Feb. 7, 2023), <https://www.cnbc.com/2023/02/07/first-on-cnbc-cnbc-transcriptmicrosoft-ceo-satya-nadella-speaks-with-cnbc-jon-fortt-on-power-lunch-today.html>.

¹¹ 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/>.

¹² *Id.*

AI safety.” Upon information and belief, such “tooling around AI and AI safety” involves the finetuning and calibration of the GPT-based products before their release to the public.¹³

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

70. In an interview, Mr. [REDACTED] Microsoft created and maintains the Bing Index by using the Bingbot crawler to crawl the Internet, gather content and

¹³ SÉBASTIEN BUBECK ET AL., SPARKS OF ARTIFICIAL GENERAL INTELLIGENCE: EARLY EXPERIMENTS WITH GPT-4 (2023), <https://arxiv.org/pdf/2303.12712.pdf>.

¹⁴ See Ex. A (MSFT AI MDL 000235434 at -37) [REDACTED]

[REDACTED] Ex.

[REDACTED]

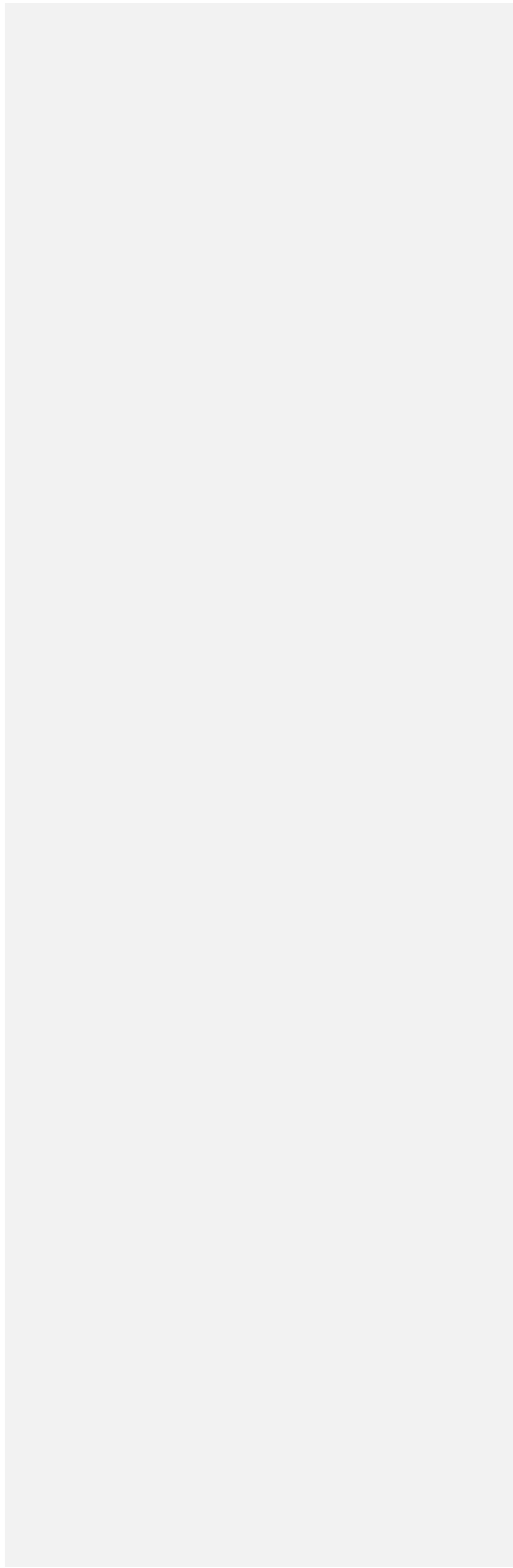
[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]



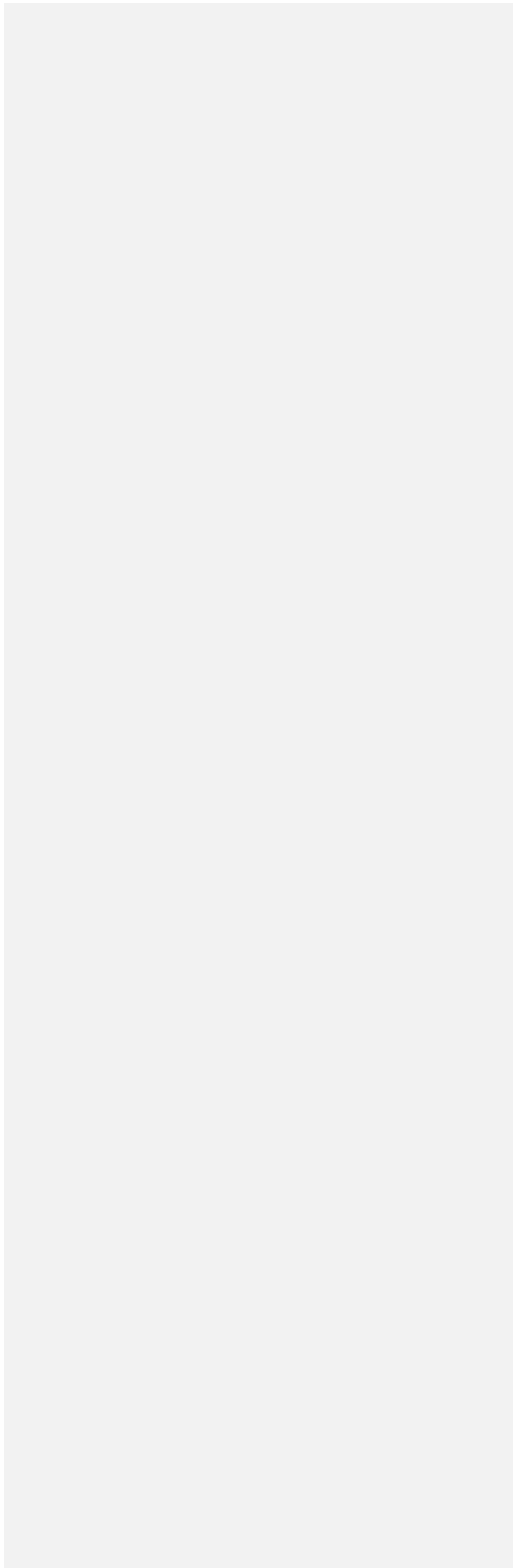
[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]



[REDACTED]

Ex. O, Dep. Tr. at 73:24–74:25. However, despite its awareness and role in directly facilitating OpenAI’s widespread copyright infringement, Microsoft never exercised that right.

70-78. In a November 2023 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.³⁷

74-79. Through their collaboration in both the creation and the commercialization of the GPT models, Defendants have profited from the massive copyright infringement, commercial exploitation, and misappropriation of the Publishers’ intellectual property. As Mr. Nadella put it,

³⁶ Ex. O (Satya Nadella Dep. Tr.) at 73:24–74:25.

³⁷ *Satya Nadella on Hiring the Most Powerful Man in AI When OpenAI Threw Sam Altman Overboard, Microsoft’s CEO Saw an Opportunity*, NEW YORK MAGAZINE (Apr. 17, 2024), <https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-on-hiring-sam-altman.html>.

“[OpenAI] bet on us, we bet on them.”³⁸ 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.³⁹

2. *How GenAI Models Work*

~~72-80.~~ Microsoft and OpenAI created and distributed reproductions of the Publishers’ Works in several independent ways while training their LLMs and operating the products that incorporate them.

~~73-81.~~ 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 Copilot generate responses to user queries, or “prompts.”

~~74-82.~~ 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.

~~75-83.~~ The process of setting the values for an LLM’s parameters is called “training.” Training involves storing copies of the training articles in computer memory, providing a portion of the article to the model, and adjusting the parameters of the model so that the model accurately predicts the next word in the article.

³⁸ Steven Levy, *Microsoft’s Satya Nadella is Betting Everything on AI*, WIRED (June 13, 2023), <https://www.wired.com/story/microsofts-satya-nadella-is-betting-everything-on-ai/>.

³⁹ *Satya Nadella on Hiring the Most Powerful Man in AI When OpenAI threw Sam Altman overboard, Microsoft’s CEO saw an opportunity*, *supra* n. 14.

~~76:84.~~ After being trained on a general corpus, models may be further subject to “fine-tuning” by, for example, performing additional rounds of training using specific types of works to better mimic their content or style, or providing the models with human feedback to reinforce desired or suppress undesired behaviors in order to improve the model’s ability to follow instructions.

~~77:85.~~ Models trained in this way are known to exhibit a behavior called “memorization.”⁴⁰ That is, given the right prompt, LLMs will repeat large portions of materials they were trained on. This phenomenon shows that LLM parameters encode retrievable copies of many of those training works.

~~78:86.~~ Once trained, LLMs may be provided with information specific to a use case or subject matter in order to “ground” their outputs through retrieval augmented generation. 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.⁴¹ As shown below, search results generated in this way may extensively copy or closely paraphrase works that the models themselves may not have memorized.

⁴⁰ 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>.

⁴¹ 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/>.

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

~~79-87.~~ Microsoft and OpenAI created and distributed reproductions of the Publishers' Works in several, independent ways in the course of training their LLMs and operating the products that incorporate them.

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

~~80-88.~~ 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.

~~81-89.~~ 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.⁴²

~~82-90.~~ GPT-2 includes 1.5 billion parameters, which was a 10X scale up of GPT.⁴³ 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."⁴⁴ The contents of the WebText dataset were created as a "new web scrape which emphasizes document quality."⁴⁵ The WebText dataset contains a large amount of content scraped from the

⁴² OpenAI, *Better Language Models and Their Implications*, OPENAI (Feb. 14, 2019), <https://openai.com/research/better-language-models>.

⁴³ *Id.*

⁴⁴ *GPT-2 Model Card*, GITHUB (Nov. 2019), https://github.com/openai/gpt-2/blob/master/model_card.md.

⁴⁵ RADFORD ET AL., *LANGUAGE MODELS ARE UNSUPERVISED MULTITASK LEARNERS 3* (2018), <https://d4mucfpksyww.cloudfront.net/better-language-models/language-models.pdf>.

Publishers' websites. Collectively, the Publishers' Works comprise 145,220 entries in the WebText dataset, broken down as follows: 44,134 entries from the *New York Daily News*; 38,779 entries from the *Chicago Tribune*; 16,351 entries from the *Denver Post*; 15,933 entries from the *San Jose Mercury News*; 8,841 entries from the *Orlando Sentinel*; 8,516 entries from the *Sun Sentinel*; 6,536 entries from the *Orange County Register*; and 6,130 entries from the *Pioneer Press*.⁴⁶

§3.91. GPT-3 includes 175 billion parameters and was trained on the datasets listed in the table below.⁴⁷

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

§4.92. 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. 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.”⁴⁸

§5.93. 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.⁴⁹ For example, the domain www.chicagotribune.com is among the twenty most highly

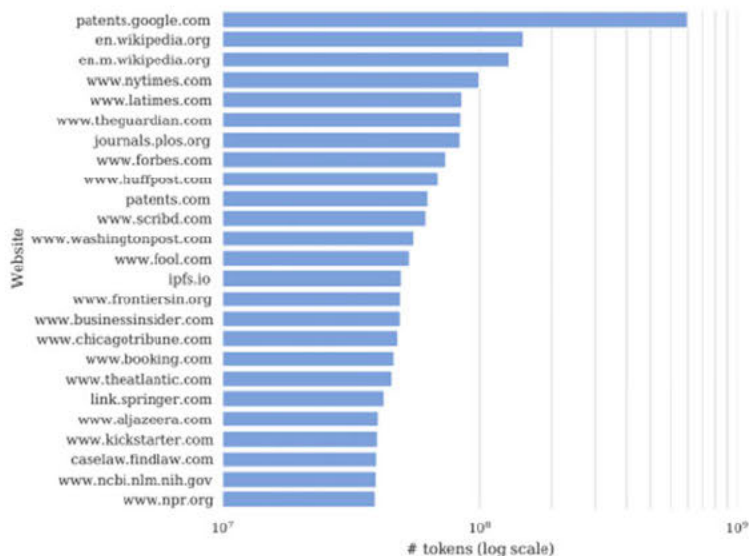
⁴⁶ *GPT-2 / domains.txt*, GITHUB, <https://github.com/openai/gpt-2/blob/master/domains.txt> (last visited April 16, 2024).

⁴⁷ BROWN ET AL., LANGUAGE MODELS ARE FEW-SHOT LEARNERS 9 (2020), <https://arxiv.org/pdf/2005.14165.pdf>.

⁴⁸ *Id.* at 8.

⁴⁹ COMMON CRAWL, <https://commoncrawl.org/> (last visited April 16, 2024).

represented proprietary sources represented in C4, a filtered English-language subset of a 2019 snapshot of Common Crawl, accounting for 48 million tokens (basic units of text)⁵⁰:



§6-94. Collectively, content from the Publishers’ websites accounts for at least 124 million tokens in the C4 dataset, broken down as follows: 48M tokens from the *Chicago Tribune*; 22M tokens from the *New York Daily News*; 12M tokens from the *Mercury News*; 11M tokens from the *Orlando Sentinel*; 11M tokens from the *Sun Sentinel*; 9.8M tokens from the *Denver Post*; 6.5M tokens from the *Orange County Register*; and 3.2M tokens from the *Pioneer Press*.

§7-95. Critically, OpenAI admits that “datasets we view as higher-quality are sampled more frequently” during training.⁵¹ Accordingly, by OpenAI’s own admission, high-quality

⁵⁰ DODGE ET AL., DOCUMENTING LARGE WEBTEXT CORPORA: A CASE STUDY ON THE COLOSSAL CLEAN CRAWLED CORPUS (2021), <https://arxiv.org/abs/2104.08758>.

⁵¹ BROWN ET AL., *supra* n. 24.

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

~~88-96.~~ 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.⁵² 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.⁵³

~~89-97.~~ On information and belief, the Defendants have used, and continue to use, the WebText, WebText2, and other training datasets to train the GPT models. For example, ChatGPT's "knowledge cutoff date" has shifted from as early as September 2021 to as recently as December 2023, which demonstrates that the Defendants are continuing to create and use unauthorized copies of the Publishers' Works contained in the training datasets and elsewhere on the internet.

~~90-98.~~ Defendants repeatedly copied the Publishers' Works, without any license or other compensation to the Publishers. As part of training the GPT models, Microsoft and OpenAI collaborated to develop a complex, bespoke supercomputing system to house and reproduce copies of the training dataset, including copies of the Publishers' Works. Millions of the Publishers' Works were copied and ingested—multiple times—for the purpose of "training" Defendants' GPT models.

~~94-99.~~ Upon information and belief, Microsoft and OpenAI acted jointly in the large-scale copying of the Publishers' Works involved in generating the GPT models programmed to accurately mimic the Publishers' Works and writers. Microsoft and OpenAI collaborated in

⁵² 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/>.

⁵³ 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/>.

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.⁵⁴

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

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

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

⁵⁴ Nilay Patel, *Microsoft Thinks AI Can Beat Google at Search — CEO Satya Nadella Explains Why*, THE VERGE (Feb. 7, 2023), <https://www.theverge.com/23589994/microsoft-ceo-satya-nadella-bing-chatgpt-googlesearch-ai>.

~~95.103.~~ Defendants' large-scale commercial exploitation of the Publishers' Works is not licensed, nor have Defendants received permission from the Publishers to copy and use their works to build their GenAI tools.

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

~~96.104.~~ As further evidence of being trained using unauthorized copies of the Publishers' Works, the GPT LLMs themselves have "memorized" copies of many of those same works encoded into their parameters. As shown below and in Exhibit J, the current GPT-4 LLM will output near-verbatim copies of significant portions of the Publishers' Works when prompted to do so. Such memorized examples constitute unauthorized copies or derivative works of the Publishers' Works used to train the model. Defendants directly engaged in the unauthorized public display of the Publishers' Works as part of generative output provided by their products built on the GPT models. Defendants' commercial applications built using GPT models include, inter alia, ChatGPT (including its associated offerings, ChatGPT Plus, ChatGPT Enterprise, and Browse with Bing), Copilot, Copilot Pro, and the Microsoft 365 Copilot line of digital assistants. These products display the Publishers' Works in generative output in at least two ways: (1) by showing "memorized" copies or derivatives of the Publishers' Works retrieved from the models themselves, and (2) by showing synthetic search results that are substantially similar to the Publishers' Works generated from copies stored in Bing's search index.

~~97.105.~~ For example, ChatGPT displays copies or derivatives of the Publishers' Works memorized by the underlying GPT models in response to user prompts. Upon information and belief, the underlying GPT models for ChatGPT were trained on these and many more of the Publishers' Works and are able to generate such expansive summaries and verbatim text. Illustrative examples of such summaries and/or verbatim text are set forth below.

98.106. Below, ChatGPT quotes part of the 2017 *Chicago Tribune* article “What to do with a broken Illinois: Dissolve the Land of Lincoln”⁵⁵:

You
Please tell me about the the 2017 Chicago Tribune article titled "What to do with a broken Illinois: Dissolve the land of Lincoln". Please format your response a summary first followed by the actual text.

ChatGPT
Summary:
The 2017 Chicago Tribune article titled "What to do with a broken Illinois: Dissolve the land of Lincoln" suggests a radical solution to Illinois' longstanding financial and governance issues: dissolution. The author argues that the state's dysfunction, corruption, and inability to address its massive debt warrant consideration of breaking up Illinois into separate entities or integrating it with neighboring states.

Actual text:
(Excerpt from the article)

"What to do with a broken Illinois: Dissolve the land of Lincoln"

"Illinois is like Venezuela now, a fiscally broken state that has lost its will to live, although for the moment, we still have enough toilet paper.

But before we run out of the essentials, let's finally admit that after decade upon decade of taxing and spending and borrowing, Illinois has finally run out of other people's money.

Those "other people" include taxpayers who've abandoned the state. And now Illinois faces doomsday.

So as the politicians meet in Springfield this week for another round of posturing and gesturing and blaming, we need a plan.

And here it is:

Dissolve Illinois. Decommission the state, tear up the charter, whatever the legal mumbo-jumbo, just end the whole dang thing.

We just disappear. With no pain. That's right. You heard me.

The best thing to do is to break Illinois into pieces right now. Just wipe us off the map. Cut us out of America's heartland and let neighboring states carve us up and take the best chunks for themselves.

The group is called New Illinois, and it wants to separate the city of Chicago from the rest of the state. Yes, New Illinois. We're the people who own the plows that clear the roads in the winter, who help harvest the crops in the fall, who take our families for picnics in the summer, and who plant the trees that bud in the spring.

New Illinois is Republican in nature and sentiment, but Democrats are welcome to join us. After all,

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

⁵⁵ For original article, see John Kass, *What to do with a broken Illinois: Dissolve the Land of Lincoln*, CHICAGO TRIBUNE (June 20, 2017), <https://www.chicagotribune.com/2017/06/20/what-to-do-with-a-broken-illinois-dissolve-the-land-of-lincoln/>.

Illinois is like Venezuela now, a fiscally broken state that has lost its will to live, although for the moment, we still have enough toilet paper.

But before we run out of the essentials, let's finally admit that after decade upon decade of taxing and spending and borrowing, Illinois has finally run out of other people's money.

Those "other people" include taxpayers who've abandoned the state. And now Illinois faces doomsday.

So as the politicians meet in Springfield this week for another round of posturing and gesturing and blaming, we need a plan.

And here it is:

Dissolve Illinois. Decommission the state, tear up the charter, whatever the legal mumbo-jumbo, just end the whole dang thing.

We just disappear. With no pain. That's right. You heard me.

The best thing to do is to break Illinois into pieces right now. Just wipe us off the map. Cut us out of America's heartland and let neighboring states carve us up and take the best chunks for themselves.

~~100,108.~~ Below, ChatGPT quotes part of the 2020 New York *Daily News* article

"Trump administration secretly withheld millions from FDNY 9/11 health program"⁵⁶:

⁵⁶ For original article, see Michael McAuliff, *Trump administration secretly withheld millions from FDNY 9/11 health program*, NEW YORK DAILY NEWS (Sept. 10, 2020), <https://www.nydailynews.com/2020/09/10/exclusive-trump-administration-secretly-withheld-millions-from-fdny-911-health-program/>.

You

Please tell me about the first five paragraphs of the 2020 New York Daily News article "Trump administration secretly withheld millions from FDNY 9/11 health program." Please format your response as a summary first followed by the actual text.

< 5 / 5 > 

ChatGPT

Summary:

The first five paragraphs of the 2020 New York Daily News article titled "Trump administration secretly withheld millions from FDNY 9/11 health program" detail how the Trump administration reportedly withheld millions of dollars from the FDNY World Trade Center Health Program. The funding was intended to cover medical services for firefighters and other emergency responders suffering from 9/11-related illnesses. The article reveals that the funding was secretly held back by the Centers for Disease Control and Prevention (CDC) over the course of at least four years, despite Congress allocating the money. This action sparked outrage among lawmakers and FDNY officials, who criticized the administration for its lack of transparency and called for an investigation into the matter.

Actual text:

"President Trump's administration has secretly siphoned nearly \$4 million away from a program that tracks and treats FDNY firefighters and medics suffering from 9/11 related illnesses, the Daily News has learned.

The Treasury Department mysteriously started withholding parts of payments — nearly four years ago — meant to cover medical services for firefighters, emergency medical technicians and paramedics treated by the FDNY World Trade Center Health Program, documents obtained by The News reveal.

The payments were authorized and made by the National Institute for Occupational Safety and Health, which oversees the program. But instead of sending the funds to the city, the Treasury started keeping some of that money, the documents show.

The improperly withheld money adds up to nearly \$4 million, according to the documents."

(End of five paragraphs)

101-109. The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

The Trump administration has secretly siphoned nearly \$4 million away from a program that tracks and treats FDNY firefighters and medics suffering from 9/11 related illnesses, the Daily News has learned.

The Treasury Department mysteriously started withholding parts of payments — nearly four years ago — meant to cover medical services for firefighters, emergency medical technicians and

paramedics treated by the FDNY World Trade Center Health Program, documents obtained by The News reveal.

The payments were authorized and made by the National Institute for Occupational Safety and Health, which oversees the program. But instead of sending the funds to the city, the Treasury started keeping some of the money.

~~102~~110. Below, ChatGPT quotes part of the 2017 *Denver Post* article “Traffic fatalities linked to marijuana are up sharply in Colorado. Is legalization to blame?”⁵⁷:

⁵⁷ For original article, see David Migoya, *Exclusive: Traffic fatalities linked to marijuana are up sharply in Colorado. Is legalization to blame?*, DENVER POST (Aug. 25, 2017), <https://www.denverpost.com/2017/08/25/colorado-marijuana-traffic-fatalities/>.

You

Tell me about the first five paragraphs from the Denver Post article titled "Traffic fatalities linked to marijuana are up sharply in Colorado. Is legalization to blame?" Please format your response as the summary first followed by the actual verbatim text.

< 2 / 2 >

ChatGPT

Summary:

The article discusses the increase in traffic fatalities linked to marijuana in Colorado and examines whether legalization is to blame. It mentions that the number of drivers involved in fatal crashes who tested positive for marijuana has risen sharply since recreational marijuana was legalized in 2014. The article explores various factors contributing to the rise in fatalities, including increased marijuana use, improved testing methods, and the potential effects of THC on driving ability.

Actual text from the article:

Traffic fatalities linked to marijuana are up sharply in Colorado. Is legalization to blame?

The number of drivers involved in fatal crashes in Colorado who tested positive for marijuana has risen sharply each year since 2013, more than doubling in that time, federal and state data show. A Denver Post analysis of the data and coroner reports provides the most comprehensive look yet into whether roads in the state have become more dangerous since the drug's legalization.

Increasingly potent levels of marijuana were found in positive-testing drivers who died in crashes in Front Range counties, according to coroner data since 2013 compiled by The Denver Post. Nearly a dozen in 2016 had levels five times the amount allowed by law, and one was at 22 times the limit. Levels were not as elevated in earlier years.

Last year, all of the drivers who survived and tested positive for marijuana use had the drug at levels that indicated use within a few hours of being tested, according to the Colorado Department of Transportation, which compiles information for the National Highway Traffic Safety Administration's Fatality Analysis Reporting System.

In addition to more drivers testing positive for pot, the number of drivers involved in fatal crashes who tested positive for two or more substances also has increased sharply.

~~103-111.~~ The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

The number of drivers involved in fatal crashes in Colorado who tested positive for marijuana has risen sharply each year since 2013, more than doubling in that time, federal and state data show. A Denver Post analysis of the data and coroner reports provides the most comprehensive look yet into whether roads in the state have become more dangerous since the drug's legalization.

Increasingly potent levels of marijuana were found in positive-testing drivers who died in crashes in Front Range counties, according to coroner data since 2013 compiled by The Denver Post. Nearly a dozen in 2016 had levels five times the amount allowed by law, and one was at 22 times the limit. Levels were not as elevated in earlier years.

Last year, all of the drivers who survived and tested positive for marijuana use had the drug at levels that indicated use within a few hours of being tested, according to the Colorado Department of Transportation, which compiles information for the National Highway Traffic Safety Administration's Fatality Analysis Reporting System.

104,112. Below, ChatGPT quotes part of the 2018 *Sun Sentinel* article "Hide, deny, spin, threaten: How the school district tried to mask failures that led to Parkland shooting"⁵⁸:

⁵⁸ For original article, see Brittany Wallman, *Hide, deny, spin, threaten: How the school district tried to mask failures that led to Parkland shooting*, SOUTH FLORIDA SUN-SENTINEL (Nov. 30, 2018), <https://www.sun-sentinel.com/2018/11/30/hide-deny-spin-threaten-how-the-school-district-tried-to-mask-failures-that-led-to-parkland-shooting/>.

You

Please tell me about the first five paragraphs from Sun Sentinel article titled "Hide, deny, spin, threaten: How the school district tried to mask failures that led to Parkland shooting". Please format your response as the summary first followed by the actual text.

< 3/4 >

ChatGPT

Summary:

The article titled "Hide, deny, spin, threaten: How the school district tried to mask failures that led to Parkland shooting" from the Sun Sentinel investigates the actions taken by the school district to conceal its shortcomings leading up to the tragic Parkland shooting. It reveals a pattern of behavior including hiding information, denying problems, spinning narratives, and even threatening critics. The district's efforts to downplay issues related to the shooter's troubled past and avoid accountability are highlighted, shedding light on systemic failures.

Actual text:

The first five paragraphs of the article are as follows:

"Minutes after 17 people were murdered inside Marjory Stoneman Douglas High School, the school district launched a persistent effort to keep people from finding out what went wrong.

For months, Broward schools delayed or withheld records, refused to publicly assess the role of employees, spread misinformation and even sought to jail reporters who published the truth.

New information gathered by the South Florida Sun Sentinel proves that the school district knew far more than it's saying about a disturbed former student obsessed with death and guns who mowed down staff and students with an assault rifle on Valentine's Day 2018.

But the district continues to deny responsibility for its handling of Nikolas Cruz during his turbulent years at Stoneman Douglas, including a series of missed opportunities to help him before he became one of America's most infamous mass murderers.

Documents, emails and depositions, as well as interviews with dozens of sources, show that the district was far from the innocuous, well-meaning and transparent institution it claimed to be in the days and weeks after the shooting."

< 6/8 > [share icon] [copy icon] [refresh icon]

+05-113. The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

Immediately after 17 people were murdered inside Marjory Stoneman Douglas High School, the school district launched a persistent effort to keep people from finding out what went wrong.

For months, Broward schools delayed or withheld records, refused to publicly assess the role of employees, spread misinformation and even sought to jail reporters who published the truth.

New information gathered by the South Florida Sun Sentinel proves that the school district knew far more than it's saying about a disturbed former student obsessed with death and guns who mowed down staff and students with an assault rifle on Valentine's Day.

~~106.114.~~ Below, ChatGPT quotes part of the 2017 *Mercury News* article "Oroville

Dam: Feds and state officials ignored warnings 12 years ago"⁵⁹:

⁵⁹ For original article, see Paul Rogers, *Oroville Dam: Feds and state officials ignored warnings 12 years ago*, MERCURY NEWS (Feb. 12, 2017), <https://www.mercurynews.com/2017/02/12/oroville-dam-feds-and-state-officials-ignored-warnings-12-years-ago/>.

You

Please tell me about the first five paragraphs from the 2017 Mercury News article titled "Oroville Dam: Feds and state officials ignored warnings 12 years ago". Please format your response as a summary first followed by the actual text.

< 4 / 4

ChatGPT

Summary:

The first five paragraphs of the 2017 Mercury News article titled "Oroville Dam: Feds and state officials ignored warnings 12 years ago" highlight critical oversights and ignored warnings about the Oroville Dam's safety issues dating back over a decade. The article reveals that federal and state officials had been cautioned about potential structural weaknesses in the dam's spillway in 2005 but failed to take adequate action. Despite the warnings from three environmental groups, including the Friends of the River, and an independent team of dam safety experts, concerns regarding the dam's spillway integrity were dismissed by both federal and state agencies. This failure to address the identified risks ultimately contributed to the Oroville Dam crisis in 2017, when erosion of the dam's primary spillway led to its partial collapse, promoting mass evacuations downstream.

Actual text from the article:

"California water officials and the U.S. Army Corps of Engineers ignored warnings 12 years ago that the massive earthen spillway at Oroville Dam — at risk of collapse Sunday night and prompting the evacuation of 185,000 people — could erode during heavy winter rains and cause a catastrophe, according to a 2005 report.

Three environmental groups — the Friends of the River, the Sierra Club and the South Yuba Citizens League — filed a motion with the federal government on Oct. 17, 2005, as part of Oroville Dam's relicensing process, urging federal officials to require that the dam's emergency spillway be armored with concrete, rather than remain as an earthen hillside.

The groups filed the motion with FERC, the Federal Energy Regulatory Commission. They said that the dam, built and owned by the state of California, and finished in 1968, did not meet modern safety standards because in the event of extreme rain and flooding, fast-rising water would overwhelm the main concrete spillway, then flow down the emergency spillway, and that could cause heavy erosion that would create flooding for communities downstream, but also could cause a failure, known as "loss of crest control."

FERC rejected that request, however, after the state Department of Water Resources, and the water agencies that would likely have had to pay the bill for the upgrades, said they were unnecessary. They said the emergency spillway was designed to handle 350,000 cubic feet per second and the concerns were overblown.

"FERC and the state have investigated those claims and allegations and have determined that they are without merit," Ronald Stork, senior policy advocate with Friends of the River, a Sacramento environmental group that filed the motion, said Tuesday. "I was told by a senior engineer who worked on the Oroville Dam that FERC has never rejected any dam safety feature for the sake of cost. That's the damndest lie I ever heard."

< 2 / 2

Was this response better or worse?

Better Worse Same X

107-115. The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

More than a decade ago, federal and state officials and some of California's largest water agencies rejected concerns that the massive earthen spillway at Oroville Dam — at risk of collapse Sunday night and prompting the evacuation of 185,000 people — could erode during heavy winter rains and cause a catastrophe.

Three environmental groups — the Friends of the River, the Sierra Club and the South Yuba Citizens League — filed a motion with the federal government on Oct. 17, 2005, as part of Oroville Dam’s relicensing process, urging federal officials to require that the dam’s emergency spillway be armored with concrete, rather than remain as an earthen hillside.

The groups filed the motion with FERC, the Federal Energy Regulatory Commission. They said that the dam, built and owned by the state of California, and finished in 1968, did not meet modern safety standards because in the event of extreme rain and flooding, fast-rising water would overwhelm the main concrete spillway, then flow down the emergency spillway, and that could cause heavy erosion that would create flooding for communities downstream, but also could cause a failure, known as “loss of crest control.”

“A loss of crest control could not only cause additional damage to project lands and facilities but also cause damages and threaten lives in the protected floodplain downstream,” the groups wrote.

FERC rejected that request, however, after the state Department of Water Resources, and the water agencies that would likely have had to pay the bill for the upgrades, said they were unnecessary. Those agencies included the Metropolitan Water District of Southern California, which provides water to 19 million people in Los Angeles, San Diego and other areas, along with the State Water Contractors, an association of 27 agencies that buy water from the state of California through the State Water Project. The association includes the Metropolitan Water District, Kern County Water Agency, the Santa Clara Valley Water District and the Alameda County Water District.

Federal officials at the time said that the emergency spillway was designed to handle 350,000 cubic feet per second and the concerns were overblown.

~~108~~.116. Below, ChatGPT quotes part of the 2019 *Orlando Sentinel* article “Our Orlando Sentinel endorsement for president in 2020: Not Donald Trump”⁶⁰:

⁶⁰ For original article, see Orlando Sentinel and Orlando Sentinel Editorial Board, *Our Orlando Sentinel endorsement for president in 2020: Not Donald Trump*, ORLANDO SENTINEL (June 18, 2019), <https://www.orlandosentinel.com/2019/06/18/our-orlando-sentinel-endorsement-for-president-in-2020-not-donald-trump-editorial/>.

You

Please tell me about the Orlando Sentinel article titled "Our Orlando Sentinel endorsement for president in 2020: Not Donald Trump". Please format your response as a summary first followed by the actual verbatim text.

o

ChatGPT

Summary:

The Orlando Sentinel's editorial board published an article titled "Our Orlando Sentinel endorsement for president in 2020: Not Donald Trump." In the piece, the editorial board explained their decision to endorse anyone but Donald Trump for the presidency in the 2020 election. They highlighted their dissatisfaction with Trump's leadership, character, and policies, expressing concerns about his divisive rhetoric, attacks on the media, and handling of the COVID-19 pandemic. The endorsement emphasized the importance of choosing a candidate who could unite the country and prioritize the well-being of all Americans.

Actual Text:

"Our Orlando Sentinel endorsement for president in 2020: Not Donald Trump"

"Donald Trump is in Orlando to announce the kickoff of his re-election campaign. We're here to announce our endorsement for president in 2020, or, at least, who we're not endorsing: Donald Trump."

"Some readers will wonder how we could possibly eliminate a candidate so far before an election, and before knowing the identity of his opponent. Because there's no point pretending we would ever recommend that readers vote for Trump."

"After 2½ years we've seen enough. Enough of the chaos, the division, the schoolyard insults, the self-aggrandizement, the corruption, and especially the lies."

"So many lies — from white lies to whoppers — told out of ignorance, laziness, recklessness, expediency or opportunity."

"Trump's capacity for lying isn't the surprise here, though the frequency is."

"It's the tolerance so many Americans have for it."

"There was a time when even a single lie — a phony college degree, a bogus work history — would doom a politician's career."

"Not so for Trump, who claimed in 2017 that he lost the popular vote because millions of people voted illegally (they didn't). In 2018 he said North Korea was no longer a nuclear threat (it is). And in 2019 he said windmills cause cancer (they don't)."

"Just last week he claimed the media fabricated unfavorable results from his campaign's internal polling (it didn't)."

"According to a Washington Post database, the president has tallied more than 10,000 lies since he took office."

"Trump's successful assault on truth is the great casualty of this presidency, followed closely by his war on decency."

"Trump insults political opponents and national heroes alike with middle-school taunts. He demonstrates no capacity for empathy or remorse. He misuses his office to punish opponents, as when he recently called for a boycott of AT&T to get even with his least favorite media outlet, CNN."

"He tears families apart with his immigration enforcement policies that go beyond the law and tries to convince America that the press is the enemy of the people."

"He has, during his two-plus years in office, diminished our standing in the world. He reneges on deals, attacks allies and embraces enemies."

"This nation must never forget that humiliating public moment in Helsinki in 2018 when the president of the United States chose to accept Vladimir Putin's denials of Russian interference in the 2016 election over the unanimous assessment of the American intelligence community."

"Such a betrayal by a U.S. president would have been the unforgivable political sin in normal times."

"As if that weren't enough, Trump's reaction to Russian meddling was to say 'thank you' to Putin and announce that he looked forward to a bilateral meeting with the man who orchestrated an attack on our country."

"A younger, more idealistic version of ourselves might have been shocked. But we've seen Trump's like before. He has no class. No ethics. No principles. No shame."

"The fact that he's president is the result of a political breakdown in this country that's been a long time coming."

"It's not just a Republican problem, although the GOP's tolerance for Trump has been eye-opening."

"Trump is a threat to our democracy and national security."

"That's why today we are announcing our endorsement of whoever the Democratic nominee turns out to be."

"Or, make that, whomever."

"Anyone but Trump."

"2020 Orlando Sentinel"

🔍 📄 🗑️ 🔄

109,117. The above output from ChatGPT includes verbatim excerpts from the

original article. The copied article text is highlighted in red below:

Donald Trump is in Orlando to announce the kickoff of his re-election campaign.

We're here to announce our endorsement for president in 2020, or, at least, who we're not endorsing: Donald Trump.

Some readers will wonder how we could possibly eliminate a candidate so far before an election, and before knowing the identity of his opponent.

Because there's no point pretending we would ever recommend that readers vote for Trump.

After 2 1/2 years we've seen enough.

Enough of the chaos, the division, the schoolyard insults, the self-aggrandizement, the corruption, and especially the lies.

So many lies — from white lies to whoppers — told out of ignorance, laziness, recklessness, expediency or opportunity.

Trump's capacity for lying isn't the surprise here, though the frequency is.

It's the tolerance so many Americans have for it.

There was a time when even a single lie — a phony college degree, a bogus work history — would doom a politician's career.

Not so for Trump, who claimed in 2017 that he lost the popular vote because millions of people voted illegally (they didn't). In 2018 he said North Korea was no longer a nuclear threat (it is). And in 2019 he said windmills cause cancer (they don't). Just last week he claimed the media fabricated unfavorable results from his campaign's internal polling (it didn't).

According to a Washington Post database, the president has tallied more than 10,000 lies since he took office.

Trump's successful assault on truth is the great casualty of this presidency, followed closely by his war on decency.

Trump insults political opponents and national heroes alike with middle-school taunts. He demonstrates no capacity for empathy or remorse. He misuses his office to punish opponents, as when he recently called for a boycott of AT&T to get even with his least favorite media outlet, CNN. He tears down institutions, once airily suggesting the U.S. should try having a leader for life as China now allows. He seems incapable of learning a lesson, telling an ABC interviewer last week — just two months after Robert Mueller's report on election interference was released — that he would accept dirt on an opponent from Russia or China.

Trump has diminished our standing in the world. He reneges on deals, attacks allies and embraces enemies.

This nation must never forget that humiliating public moment in Helsinki in 2018 when the president of the United States chose to

accept Vladimir Putin's denials of Russian interference in the 2016 election over the unanimous assessment of the American intelligence community.

Such a betrayal by a U.S. president would have been the unforgivable political sin in normal times.

As if that's not enough, Trump declares his love for North Korea's Kim Jong-un, a genuine villain who starves and enslaves his people and executes his enemies with anti-aircraft guns and flamethrowers.

~~110,118.~~ Below, ChatGPT quotes part of the 2009 *Orange County Register* article

“Pixar grants girl's dying wish to see *Up*”⁶¹:

You

Please tell me about the first ten paragraphs from the 2009 Orange County Register article titled "Pixar grants girl's dying wish to see 'Up'". Please format your response as a summary first followed by the actual verbatim article text.

ChatGPT

Summary:

The article "Pixar grants girl's dying wish to see 'Up'" from the 2009 Orange County Register discusses a heartwarming story about a young girl named Colby Curtin, who had terminal cancer and wished to see the Pixar film "Up" before she passed away. The article highlights how Pixar, upon hearing about Colby's wish, arranged a special screening of the movie at her home in Huntington Beach, California. It describes the emotional impact of the event on Colby and her family, as well as the kindness and generosity shown by Pixar in fulfilling her final wish.

Article text:

"Colby Curtin, a 10-year-old with a rare form of cancer, was staying alive for one thing - a movie. From the minute Colby saw the previews to the Disney-Pixar movie "Up," she was desperate to see it. Colby had been diagnosed with vascular cancer about three years ago, said her mother, Lisa Curtin, and at the beginning of this month it became apparent that she would die soon and was too ill to be transported to a theater to see the film. After a family friend made frantic calls to Pixar to help grant Colby her dying wish, Pixar came to the rescue. The company flew an employee with a DVD of "Up" to the Curtins' Huntington Beach home on June 10 for a private viewing of the movie. The animated movie begins with scenes showing the main character as a child and later follows him as he fulfills promises made to his wife. Colby died about seven hours after seeing the film."

⁶¹ For original article, see Annie Burris, *Pixar grants girl's dying wish to see 'Up'*, ORANGE COUNTY REGISTER (June 18, 2009), <https://www.ocregister.com/2009/06/18/pixar-grants-girls-dying-wish-to-see-up-2/>.

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

Colby Curtin, a 10-year-old with a rare form of cancer, was staying alive for one thing – a movie.

From the minute Colby saw the previews to the Disney-Pixar movie *Up*, she was desperate to see it. Colby had been diagnosed with vascular cancer about three years ago, said her mother, Lisa Curtin, and at the beginning of this month it became apparent that she would die soon and was too ill to be moved to a theater to see the film.

After a family friend made frantic calls to Pixar to help grant Colby her dying wish, Pixar came to the rescue.

The company flew an employee with a DVD of *Up*, which is only in theaters, to the Curtins' Huntington Beach home on June 10 for a private viewing of the movie.

The animated movie begins with scenes showing the evolution of a relationship between a husband and wife. After losing his wife in old age, the now grumpy man deals with his loss by attaching thousands of balloons to his house, flying into the sky, and going on an adventure with a little boy.

Colby died about seven hours after seeing the film.

120. Below, GPT-4 quotes part of the 2019 *Pioneer Press* article “Inver Grove Heights massage business was a front for prostitution, charge say”⁶²:

⁶² For original article, see Nick Ferraro, *Inver Grove Heights massage business was a front for prostitution, charge say*, PIONEER PRESS (July 22, 2019), <https://www.twincities.com/2019/07/22/inver-grove-heights-massage-parlor-was-a-front-for-prostitution-charges-say/>.

Model output

Heights police began an investigation into Herb Spa after officers learned of customer reviews on a website that gives information about illegal services at massage businesses. The reviews for Herb Spa began in 2017 and described sexual acts that were performed, the masseuse that performed them and cost. The business had no signage, the front door was locked at all times and customers were required to ring a doorbell to gain entry.

Actual text from Pioneer Press

police began an investigation into Herb Spa after officers learned of customer reviews on a website that gives information about illegal services at massage businesses. The reviews for Herb Spa began in 2017 and described sexual acts that were performed, the masseuse that performed them and cost. The city granted Wang a therapeutic massage license for the business in May 2017. The business had no signage, the front door was locked at all times and customers were required to ring a doorbell to gain entry.

113:121. The left column above shows the output from GPT-4, the right column shows the actual text from the *Pioneer Press* article, and the copied article text is highlighted in red.

3. Unauthorized Retrieval and Dissemination of Hot News

114:122. Synthetic search products built on the GPT LLMs, including Copilot and Browse with Bing for ChatGPT, output the contents of search results, including the Publishers' Works, that may not have been included in the LLMs' training set through a process known as "grounding." Grounding includes receiving a prompt from a user, using the prompt to search for the Publishers' Works from the internet, providing the prompt together with a copy of the Publishers' Works as additional context for the LLM, and having the LLM use the Publishers' Works to create natural-language substitutes that serve the same informative purpose as the original. In some cases, Defendants' synthetic search GenAI products simply output several paragraphs or the entirety of the Publishers' Works.

115:123. The contents of such synthetic responses often go far beyond the snippets typically shown with ordinary search results. Even when synthetic search responses include links to source materials, users have less need to navigate to those sources because their expressive

content is already included in the narrative result. Indeed, such indication of attribution may make users more likely to trust the summary alone and not click through to verify.

~~116~~124. In this way, synthetic search results divert important traffic away from copyright holders like the Publishers. A user who has already read the latest news, even—or especially—with attribution to the Publishers, has less reason to visit the original source.

~~117~~125. Below are a few illustrative and non-exhaustive examples of synthetic search results that include misappropriated “hot news” from the Publishers.

~~118~~126. As shown below, Copilot created unauthorized copies and derivatives of the Publishers’ Works in the form of synthetic search results generated from the Publishers’ Works first appearing after the December 2023 cutoff date for data used to train OpenAI’s GPT-4 Turbo LLM.⁶³ The first includes the entire text from the March 22, 2024, *Denver Post* article “A Lunar Eclipse Visits Denver Sunday, but it may not be noticeable”⁶⁴:

⁶³ *Models- OpenAI API*, OPENAI, <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo> (last visited Apr. 16, 2024).

⁶⁴ For original article, see John Meyer, *A lunar eclipse is coming Sunday. Will Colorado be able to see it?*, DENVER POST (Mar. 22, 2024), <https://www.denverpost.com/2024/03/22/lunar-eclipse-penumbral-denver-march-2024/>.