UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: 1:23-cv-11195-SHS-OTW 1:24-cv-04872-SHS-OTW 1:24-cv-03285-SHS-OTW 1:24-cv-01515-SHS-OTW 1:25-cv-04315-SHS-OTW	25-md-3143 (SHS) (OTW) Hon. Sidney H. Stein Hon. Ona T. Wang ORAL ARGUMENT REQUESTED REDACTED — PUBLIC VERSION

DEFENDANTS' MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT OF MOTION FOR SUMMARY JUDGMENT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

CASES

STATUTES

RULES

OTHER AUTHORITIES

I. INTRODUCTION[1]

This motion concerns two technologies. The first is the large language model, trained on a vast and diverse array of material, including news. The second is Browse, OpenAI's automated web browsing service, which performs research in response to user requests. In both cases, the model draws on what it has read to answer a user's questions in the model's own words.

News Plaintiffs claim that doing so infringes their copyrights. But News Plaintiffs have no right to prevent OpenAI from telling its users that Gustavo Dudamel has been appointed music director of the New York Philharmonic, or that the Dow hit a 52-week high of 54,744 on August 5, 2026, or that the battery on the new iPhone lasts for 24 hours and 31 minutes. Those are facts, like millions of others that News Plaintiffs publish.

Nobody owns facts, just as nobody owns language, grammar, or style. That is because "every idea, theory, and fact in a copyrighted work becomes instantly available for public exploitation at the moment of publication." *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003). Those who expend sweat or shoe-leather to report the facts do not gain a monopoly on the truths they report, because "the primary objective of copyright is not to reward the labor of authors, but to promote the Progress of Science and useful Arts." *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349 (1991). The technologies at issue here serve that objective directly, and their public benefits are vast.

Because they are trained on such a broad array of material, OpenAI's models have read about how the Giants won the pennant in 1951, and can tell you about it in their own words, however you like—in terse bullet points, or matter-of-fact prose, or anecdotes from the bleachers at the Polo Grounds. Because Browse lets those models access the web, a user can have

[1] In all quotations appearing herein, all internal quotation marks, citations, alterations, and the like have been omitted, and all emphases added, unless otherwise noted.

ChatGPT go out and read fifty sources on hurricanes or headphones or House hearings and get back a synthesis of the key facts, complete with links to their sources. "The ultimate test of fair use is whether the copyright law's goal of promoting the Progress of Science and useful Arts would be better served by allowing the use than by preventing it." *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 608 (2d Cir. 2006). Here, the answer is plain. These technologies put the world's facts within reach of anyone who asks, in whatever form is most useful to them. That is progress, and copyright law does not stand in its way.

II. BACKGROUND

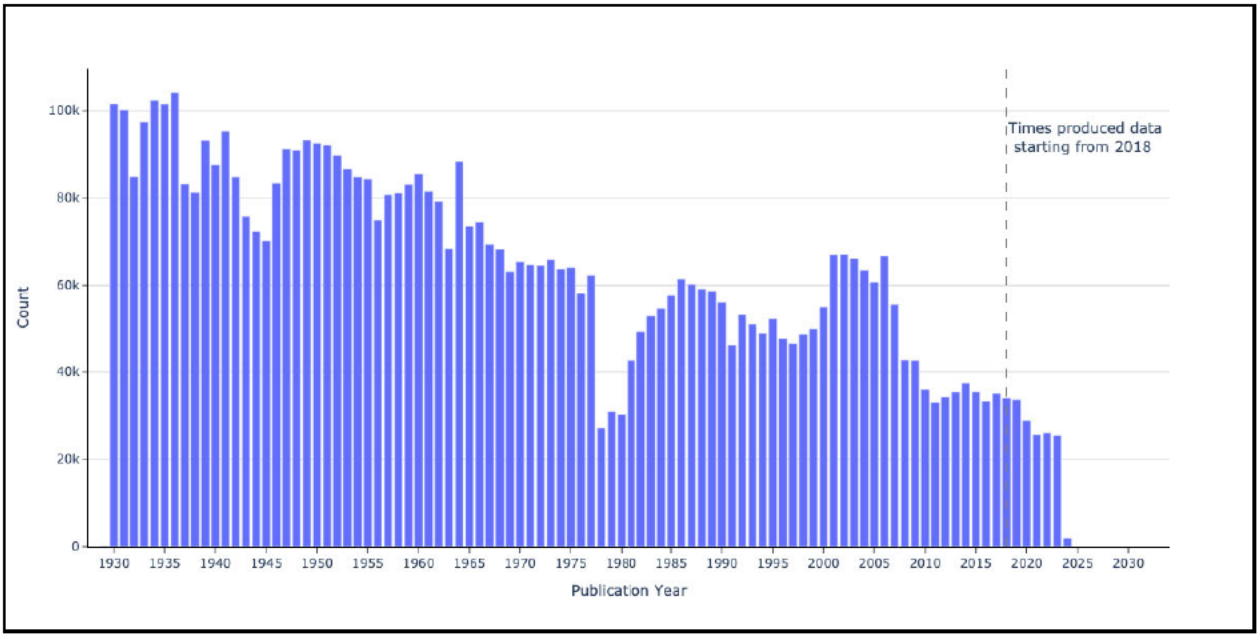
A. Technical background relevant to pretraining

The Books MSJ describes the process by which OpenAI trained its LLMs and associated measures OpenAI takes to prevent "regurgitation" of copyrighted content. *See* Books MSJ Section II(D). The same practices apply to the categories of training data at issue in this brief as well. SUF ¶¶ 1–81.

For pretraining, OpenAI focused on collecting a broad and diverse swath of text from many sources on the Internet, rather than targeting any particular category of content (*e.g.*, news articles). SUF ¶¶ 23, 30. That is because OpenAI sought to create "generalist" LLMs that could "seamlessly mix together or switch between many tasks and skills" with "fluidity and generality," rather than excel at only particular tasks. SUF ¶ 138. To do so, OpenAI used a filtered version of an enormous, publicly available index of billions of webpages created by the non-profit organization Common Crawl. SUF ¶ 139. Some of those webpages contained news articles. SUF ¶ 140. OpenAI also built new datasets that contained text from a wide variety of other websites, some of which included news articles. SUF ¶ 141.

B. News Plaintiffs, the Asserted Works, and the allegations in the News Cases

News Plaintiffs are news or news-related organizations. They include: (1) the New York Times ("The Times"), (2) eight regional newspapers ("The Daily News Plaintiffs" or "DNP"), (3) digital media conglomerate Ziff Davis, (4) The Center for Investigative Reporting ("CIR"), and (5) The Intercept. Collectively, News Plaintiffs assert infringement of millions of news articles ("the Asserted Works").[2] All of the Asserted Works are published. Their publication dates vary dramatically, however, spanning from 1928 to 2024. SUF ¶¶ 207–11, 213. This chart, for example, shows the age distribution for the 6,030,928 articles asserted by The Times:



SUF ¶¶ 207–08.

With respect to those Asserted Works—at least some of which OpenAI assumes for the purposes of this motion were used to pretrain the LLMs at issue—News Plaintiffs have alleged:

[2] News Plaintiffs have identified over 10.8 million asserted works in this case. SUF ¶ 203. But according to their own experts only 6.3 million of those works were used to train OpenAI's LLMs. SUF ¶ 204. News Plaintiffs have presented no evidence of copying for the remaining works and therefore summary judgment should be granted as to those approximately 4.5 million works. Fed. R. Civ. P. 56(a); *Jorgensen v. Epic/Sony Recs.*, 351 F.3d 46, 50 (2d Cir. 2003) (granting summary judgment given "the absence of material evidence supporting an essential element of [plaintiff's] copyright infringement claim").

- ChatGPT regurgitates "verbatim" or "closely summariz[es]" them. *See, e.g.*, The Times Third Am. Compl. ¶¶ 4–5, Dkt. No. 1677;
- OpenAI's alleged use has caused a decrease in traffic to their websites. *See, e.g.*, *id.* ¶¶ 110, 136.

The undisputed record evidence, however, tells a different story.

Regurgitation. As explained in the Books MSJ, OpenAI's LLMs are not designed to "regurgitate" their training data, and OpenAI has developed and implemented safeguards to prevent it from occurring. Books MSJ Section II(D). Accordingly, when OpenAI's expert searched for Asserted Works in a sample of 20 million ChatGPT conversation logs produced in this case, he found only 24 instances of verbatim regurgitation—*i.e.*, an alleged regurgitation rate of 0.00012%. SUF ¶ 219. News Plaintiffs' experts ran several different analyses, all of which showed similarly minute regurgitation rates: 0.00011% from The Times/DNP (SUF ¶ 220), 0.000008% from CIR (SUF ¶ 222), 0.000002% from The Intercept (SUF ¶ 221), and 0.00002% from Ziff Davis (SUF ¶ 224). The longest identified verbatim regurgitations were 29 words (SUF ¶ 223) and 43 words (SUF ¶ 225).

In addition to analyzing actual user chats, OpenAI's expert conducted two different studies on the at-issue models. In one study, he submitted prompts about 710 randomly selected Asserted Works and found that in over 68,000 GPT model outputs, there were ***zero*** instances of regurgitation. SUF ¶ 226. In the other study, he used "adversarial" prompting techniques targeting 19 randomly sampled works (one from each of the 19 outlets across the five News Plaintiffs). SUF ¶ 227. He submitted over 9,400 prompts to attempt to induce regurgitation of the Asserted Works. SUF ¶ 228. Even with these extreme techniques no ordinary user would conceive of, let alone deploy, ChatGPT regurgitated on average only 2.3% of the content from the selected works. SUF ¶ 228. Experts for The Times, DNP, and Ziff Davis performed their own

"forced" regurgitation studies. They used over 250 million prompts, non-public GPT models, and specialized model configurations available only through the API interface (not the consumer-facing ChatGPT). SUF ¶¶ 229–30; 234–35. Even then, the experts' extreme attempts triggered partial regurgitations of less than 2.6% of those Plaintiffs' Asserted Works. SUF ¶ 239. Within that small subset of works, they extracted on average under 4.4% of the targeted work. SUF ¶¶ 233, 238.

Web Traffic. The evidence shows that News Plaintiffs' web traffic has not been negatively impacted by ChatGPT. As News Plaintiffs' own expert admits, since ChatGPT's release, "the traffic for the New York Times [redacted]. Traffic for the Daily News Plaintiffs [redacted]. And the traffic for Ziff Davis [redacted]." SUF ¶ 240. Dr. Avi Goldfarb, an expert economist retained by OpenAI, conducted a detailed analysis of individual user data to assess how, if at all, users' adoption of ChatGPT impacted visits to the News Plaintiffs' websites. He found no impact.[3] SUF ¶¶ 241–43. News Plaintiffs' experts critique Dr. Goldfarb's analysis, but they have presented no empirical analysis of their own to show that ChatGPT has caused a decline in traffic to News Plaintiffs' websites. SUF ¶¶ 247–49, 253, 258–64.

Nor has ChatGPT negatively impacted News Plaintiffs' financial performance. The Times's revenues have increased substantially since ChatGPT was released. SUF ¶ 275. The Times has also hit company-record highs in subscribers, surpassing its goal of 12 million subscribers in 2025, and is on pace to meet its stated goal of reaching 15 million subscribers by the end of 2027. SUF ¶ 276. DNP's digital subscriptions [redacted]. SUF ¶ 278. Ziff Davis generated a 3.5% growth in revenue in 2025 and, as of at least the

[3] Dr. Goldfarb did not conduct a user-level data analysis for CIR or The Intercept because there was insufficient data to do so. Instead, he analyzed their web traffic data and found no impact on them, either. SUF ¶¶ 244–45.

fiscal year ended 2024, Ziff Davis did not believe that the threats posed to generative AI and related technologies had been material to its businesses to date. SUF ¶¶ 279–80. It also declined to identify generative AI as a material risk in its 2025 10-K, despite admitting that [redacted] it would have been "[redacted]" SUF ¶ 281. The Intercept's annual individual member contributions post-ChatGPT are [redacted] than they were pre-ChatGPT, SUF ¶¶ 283–84, and following its 2024 merger, CIR's [redacted] resulting in a [redacted]. SUF ¶¶ 286–87.

Generative AI and Journalism. As a Times executive put it in August 2025, "[redacted]" SUF ¶ 331. Indeed, News Plaintiffs have made extensive use of OpenAI's LLMs. SUF ¶¶ 289–306, 309–15, 317–30. For example, The Times uses OpenAI's LLMs through a tool called "ChatExplorer." SUF ¶¶ 317–26; *see also* Def.'s Mem., Dkt. No. 639. The Times used ChatExplorer to [redacted]. SUF ¶¶ 318–25. The Times also uses OpenAI's LLMs to help with advertising and branding campaigns, and The Times's witnesses admitted that this use "[redacted]." SUF ¶¶ 327–28. Daily News Plaintiffs used OpenAI's LLMs to [redacted]. *See e.g.,* SUF ¶¶ 309, 326. And one of their executives testified that he "[redacted]" and "[redacted]" SUF ¶ 335.

III. LEGAL STANDARD & FAIR USE PRINCIPLES

The legal standard and key fair use principles are set forth in the Books MSJ at Sections III–IV.

IV. ARGUMENT

A. OpenAI's alleged use of the Asserted Works for pretraining is fair use.

OpenAI's alleged use of News Plaintiffs' Asserted Works to pretrain its LLMs is fair use because the purpose is highly transformative, the works are highly factual, and pretraining has caused no cognizable harm to the market for or value of those works.

Factor 1: Any use of News Plaintiffs' works by OpenAI for pretraining is highly transformative. News Plaintiffs wrote their works to engage human readers. SUF ¶ 205. OpenAI's alleged pretraining use was for an entirely different purpose: to distill broad patterns of human language and facts about the world from an enormous text corpus, with the ultimate goal of developing LLMs that could effectively understand human language and generate entirely new text. This use is "among the most transformative many of us will see in our lifetimes." *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1033 (N.D. Cal. 2025).

This use was "justified" because it "rendered a valuable service to the public . . . without allowing public access to the copy." *Romanova v. Amilus Inc.*, 138 F.4th 104, 115 (2d Cir. 2025); *see also Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 531–32 (2023). Over a billion people use OpenAI's LLMs for a wide variety of uses. SUF ¶¶ 21–22. News Plaintiffs themselves make extensive use of OpenAI's LLMs. SUF ¶¶ 289–307, 309–15, 317–30. There is simply "no serious question" that factor one favors fair use. *See Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1044 (N.D. Cal. 2025); *see also* Books MSJ at Section V(A).

News Plaintiffs' arguments to the contrary fail. They allege that ChatGPT extensively regurgitates the Asserted Works, but their own experts' analysis proves otherwise: regurgitation is extremely rare, with rates ranging from 0.000002% to 0.00011%. *See supra* Section II(B). People do not use ChatGPT to reproduce verbatim copies of old news articles that OpenAI's LLMs were pretrained on; indeed, OpenAI's models are trained to refuse such requests.[4]

News Plaintiffs suggest that pretraining is not transformative because OpenAI's LLMs can be used to generate news content (in addition to many other types of content). This argument fails because News Plaintiffs do not own the facts embodied in the articles they write. "[E]very idea, theory, and fact in a copyrighted work becomes instantly available for public exploitation at the moment of publication." *Eldred*, 537 U.S. at 219. Indeed, this Court has already rejected News Plaintiffs' claim that OpenAI's LLMs infringe their works by summarizing the facts in them. Op. at 41, *New York Times v. Microsoft*, No. 1:23-cv-11195-SHS-OTW (S.D.N.Y. 2025), Dkt. No. 514 ("Works are not substantially similar as a matter of law when their only similarity concerns general or underlying facts and ideas, which are not copyrightable."). Using copyrighted content to create a tool that makes such facts more easily discoverable affirmatively *favors* fair use. *See Authors Guild v. Google, Inc.*, 804 F.3d 202, 217 (2d Cir. 2015) (*Google Books*).

Factor 2: The Asserted Works are published, highly factual works, which favors fair use. *See* Books MSJ at Section V(B); S*tewart v. Abend*, 495 U.S. 207, 237 (1990); *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 455 n.40 (1984); 4 Patry on Copyright § 10:138 ("Newspapers . . . should be subject to liberal appropriation since they typically contain little expression."). Many have thin-to-no copyright protection, like the "Sports Results" from July 13,

[4] To the extent News Plaintiffs assert that any particular output from OpenAI's LLMs is substantially similar to an Asserted Work, that assertion would require a separate analysis of copyright infringement.

1941, a "TV Schedule" for the 1976 Winter Olympics, and the "Colorado snow totals for Oct. 19, 2023." SUF ¶¶ 217–18.

Factor 3: The amount of any use by OpenAI was reasonable in relation to its transformative pretraining purpose. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586-87 (1994); *see also Bartz*, 787 F. Supp. 3d at 1030 (copying entire books for pretraining was reasonable); *Kadrey*, 788 F. Supp. 3d at 1050 (same). That entire articles may have been used in pretraining does not change the calculus—what matters is that OpenAI's LLMs were designed to not display protected content from copyrighted works. That also favors fair use. *See Google Books*, 804 F.3d at 222; Books MSJ at Section V(C).

Factor 4: OpenAI's alleged use is not a "substitute" for the protected expression in the Asserted Works because ChatGPT is designed not to display that expression to users, but to generate new text. *See Google Books*, 804 F.3d at 223; *see also infra* Section IV(B)(2).

Rare and isolated instances of regurgitation do not change the analysis. In *Google Books*, the Court recognized that Google's intentional display of small portions of copyrighted text could cause some lost book sales but held that "some loss of sales does not suffice to make the copy an effectively competing substitute that would tilt the weighty fourth factor in favor of the rights holder in the original. There must be a ***meaningful*** or ***significant*** effect upon the potential market for or value of the copyrighted work." *Google Books*, 804 F.3d at 224. Here, OpenAI does not intentionally display any copyrighted expression to ChatGPT users, and the amount of regurgitated text that News Plaintiffs' experts managed to force out of OpenAI's LLMs for any one work is on average substantially lower than the amount the plaintiffs' experts in *Google Books* were able to extract from a book. *Compare supra* Section II(B) (4.4%) *with Google Books*, 804 F.3d at 224 (16%). There is no evidence of any impact on News Plaintiffs, let alone,

a meaningful or significant impact.

News Plaintiffs have failed to assert a legally cognizable theory of market harm. Their experts opine that ChatGPT "substitutes" for (unidentified) "news content" because it "satisf[ies] user demand for information." SUF ¶ 206. This precise argument was rejected in *Google Books*, which held that lost sales traceable to the provision of unprotected facts from original works are not the type of market harm that copyright law condemns. 804 F.3d at 224; *accord Wright v. Warner Books, Inc.*, 953 F.2d 731, 739 (2d Cir. 1991); *NXIVM Corp. v. Ross Institute*, 364 F.3d 471, 482 (2d Cir. 2004) ("[T]he relevant market effect with which we are concerned is the market for plaintiffs' 'expression,' and thus it is the effect of defendants' use of that expression on plaintiffs' market that matters, not the effect of defendants' work as a whole.").

Even if News Plaintiffs' theory of harm were cognizable (and it is not), it still would fail because News Plaintiffs have not been harmed. The Times has more revenues, subscribers, and [redacted] than ever before. SUF ¶¶ 275–77. Other News Plaintiffs have experienced [redacted] in pre-ChatGPT traffic trends, and/or their financial performance has improved since ChatGPT was launched in November 2022. *See supra* Section II(B). Moreover, OpenAI's expert analyzed individual user data and concluded that user-adoption of ChatGPT had no negative impact on News Plaintiffs' web-traffic. *Id.* News Plaintiffs offer conclusory statements from their experts, but no corresponding analysis of their own.

This lack of impact is unsurprising. The Asserted Works are, literally, old news. In the case of The Times, over 95% of them were published before OpenAI was even founded in November 2015, and two-thirds were published before 1981. SUF ¶ 208. Even if it were true that users were turning to ChatGPT to learn the facts contained in these old articles (and there is no evidence this is happening), that would have zero impact on demand for the present-day breaking

news The Times specializes in. This is true even for articles that are much more recent. By the time an LLM pretrained on those articles is made available to the public, those articles are at least several months old. SUF ¶¶ 142, 208–13.

Unable to show actual harm, News Plaintiffs resort to speculation that web traffic or sales will decrease in the future. SUF Section IX(C). This litigation-driven position contradicts their own public statements. The Times says that generative AI "[redacted]" SUF ¶ 331; *see also* SUF ¶ 332 ("[redacted]"); SUF ¶¶ 333–34. Regardless, speculation about future harm—even if it were cognizable—is insufficient to avoid summary judgment. *See, e.g.*, *Fujitsu Ltd. v. Fed. Exp. Corp.*, 247 F.3d 423, 428 (2d Cir. 2001) (nonmoving party "may not rely on conclusory allegations or unsubstantiated speculation").

Finally, News Plaintiffs make the same licensing-fee argument as Books Plaintiffs: that their fourth-factor harm is OpenAI's failure to pay them for the use in question in this case. That argument fails in the first instance because News Plaintiffs are not entitled to a license fee for OpenAI's transformative use. *See Bill Graham Archives*, 448 F.3d at 614. But regardless, the only two courts to confront this circular argument have rejected it. *See Kadrey*, 788 F. Supp. 3d at 1052 (rejecting argument that failure to license for LLM training constituted cognizable harm in order "to prevent the fourth factor analysis from becoming circular and favoring the rightsholder in every case"); *Bartz*, 787 F. Supp. 3d at 1032 ("[S]uch a market for that use is not one the Copyright Act entitles Authors to exploit").[5]

* * *

[5] OpenAI expects that News Plaintiffs will invoke certain commercial partnerships that OpenAI has entered to argue that there is a market for LLM training licenses that has been harmed by OpenAI's failure to license their works. Nothing about that argument can or does change the analysis presented here, and OpenAI will address it in due course.

Using publicly available content from the Internet to pretrain an LLM is fair use.

B. OpenAI's alleged use of the Asserted Works in Browse is also permissible.

News Plaintiffs also allege OpenAI created copies of some of the Asserted Works through its automated web browsing service ("Browse") in ChatGPT, which implicates copying that is separate from the pretraining process for OpenAI's LLMs. Because OpenAI's alleged use of the Asserted Works in Browse is for purposes of extracting factual content, such copying is a fair use as a matter of law, *see* 17 U.S.C. § 107, or, at the very least, impliedly licensed.

1. Technical background relevant to Browse

When ChatGPT launched in November 2022, people around the world began to use it to "write jokes (some of which are actually funny)," write "working computer code," and "explain scientific concepts at multiple levels of difficulty."[6] While ChatGPT excelled at generating new text, the substance of its responses was limited to information gleaned from its training data, which meant that there was a "knowledge cutoff" after which point the tool was blind to the world. SUF ¶ 142.

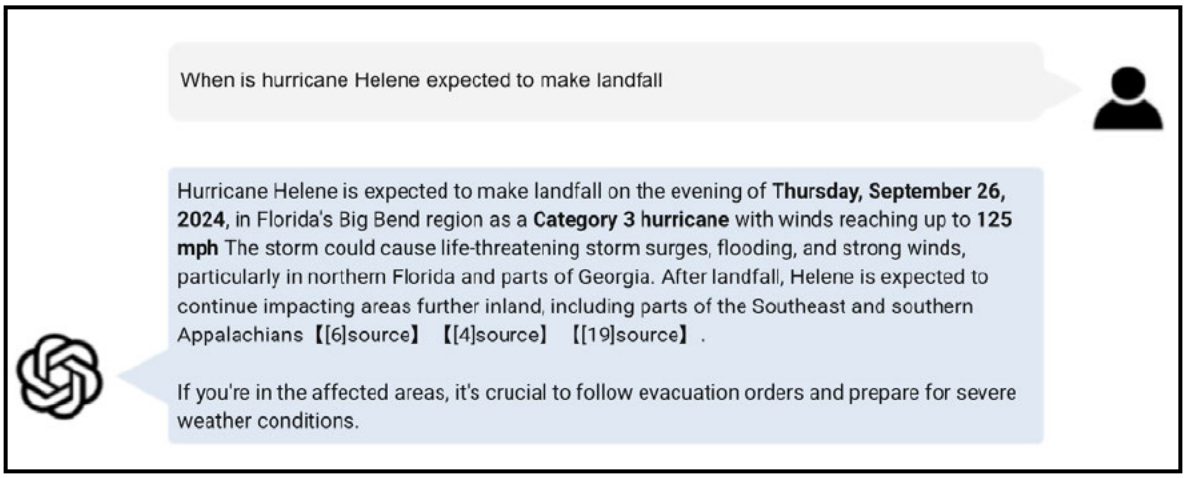
To address that limitation, OpenAI began exploring ways to enable ChatGPT to "browse" the Internet for current information when necessary to inform a response. SUF ¶ 143. In early 2023, OpenAI implemented a functionality ("Browse"[7]) that mirrored the way in which a human might look something up on the Internet. SUF ¶ 143. When a user asked ChatGPT a question (*e.g.*, "*when does MoMA open today*"), ChatGPT would convert that prompt into a search query, adding contextual information (*e.g.*, "*MoMA NYC opening hours July 23, 2026*"). SUF ¶ 145.

---

[6] Kevin Roose, *The Brilliance and Weirdness of ChatGPT*, N.Y. TIMES (Dec. 5, 2022), https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html.

[7] OpenAI's web-browsing service has been known at various times as "Browse with Bing," "Browse," or "ChatGPT Search." Plaintiffs refer to it as "Retrieval Augmented Generation" or "RAG," which refer to a process in which a LLM retrieves information from external sources, including webpages, by using search engines and web-retrieval techniques. *See* SUF ¶ 144.

ChatGPT would send that query to Microsoft Bing, which returned a list of results with links to relevant webpages, along with short "snippets" of text, similar to what a user would see if they ran the search on www.Bing.com.[8] SUF ¶¶ 146, 148. ChatGPT analyzed the search results returned by Bing to determine which webpages might contain relevant information, just as a user may scan the search results, and then ChatGPT sent an automated request to websites for the content of those webpages.[9] SUF ¶ 150. After receiving the content, ChatGPT would extract the relevant factual information and synthesize it into a response to the user. SUF ¶ 154. Importantly, OpenAI designed ChatGPT to share *facts* from websites, not the *creative expression* they contain. SUF ¶¶ 155–56.[10] ChatGPT also displayed links to the websites or "sources" used to inform its response so that a user could further explore the topic or verify ChatGPT's response. SUF ¶ 157. Here is how it looked in a real user chat from September 26, 2024:



SUF ¶ 158.

[8] Website operators can opt out of having their websites indexed by Bing. SUF ¶ 147.

[9] In some instances, ChatGPT used the search results returned by Bing to inform its response to the user without accessing the associated webpages. SUF ¶ 151.

[10] Soon after OpenAI launched Browse, OpenAI "learned that the browsing beta can occasionally display content in ways we don't want, e.g. if a user specifically asks for a URL's full text, it may inadvertently fulfill this request." SUF ¶ 159. OpenAI "temporarily disabl[ed] Browse" while it fixed this issue. SUF ¶ 160. OpenAI fixed the issue and re-launched Browse several months later. SUF ¶¶ 160–61.

OpenAI designed Browse to comply with the widely known and used industry standard Internet protocol known as the Robots Exclusion Protocol or "robots.txt." SUF ¶ 166. That protocol allows website owners to express their desire to "opt-out" of receiving automated requests for webpage content from specific "user agents" commonly referred to as "bots." SUF ¶ 167. News Plaintiffs have known about and used the robots.txt protocol for years. SUF ¶¶ 177–82.

In its March 2023 blog post announcing Browse, OpenAI disclosed that its user-agent was "ChatGPT-User," and that to "adhere to the web's norms," it was "configured to honor websites' robots.txt files." SUF ¶ 183. Websites that preferred to "opt out" of requests from "ChatGPT-User" could insert this simple text string into their existing robots.txt file:

```
User-agent: ChatGPT-User
Disallow: /
```

SUF ¶ 184.[11]

OpenAI's blog post was widely reported and well understood by prominent news publishers, including News Plaintiffs.[12] For example, on April 1, 2023, just nine days after OpenAI's announcement, The Times employees discussed whether they should "[redacted]." SUF ¶¶ 186–87. A few days later, The Times confirmed that "[redacted]." SUF ¶ 188. Ziff Davis documents show that [redacted]

[11] OpenAI also began using the user agent "OAI-SearchBot" for Browse in 2024. SUF ¶ 185.

[12] *See also* Richard Fletcher, *How many news websites block AI crawlers?*, REUTERS INST. (Feb. 22, 2024), https://reutersinstitute.politics.ox.ac.uk/how-many-news-websites-block-ai-crawlers ("By the end of 2023, 48% of the most widely used news websites across ten countries were blocking OpenAI's crawlers.").

[redacted]. SUF ¶ 189. The other News Plaintiffs also actively managed their robots.txt files and knew or had reason to know about OpenAI's bots. SUF ¶¶ 177–82, 186.

Despite this knowledge, News Plaintiffs did nothing. Instead, The Times waited nearly a year before it blocked OpenAI's user agent, and Ziff Davis waited [redacted] before it did so. *See* SUF ¶¶ 190–91. Other Plaintiffs did not block OpenAI's user agent until [redacted] (DNPs), January 2024 (Intercept), or after [redacted] (CIR). SUF ¶¶ 192–94.

Websites could also opt-out by configuring their servers to "hard block" requests from OpenAI, and OpenAI published the IP addresses used by its bots to enable easy blocking. SUF ¶¶ 195–96. While The New York Times implemented a hard block for ChatGPT-User in April 2024, CIR, The Intercept, and the Daily News Plaintiffs [redacted] and Ziff Davis [redacted]. SUF ¶¶ 197–202.

2. Browse's use of the Asserted Works to extract factual information constitutes fair use.

a) Factor 1: OpenAI's purpose for any use in connection with Browse favors fair use.

Browse serves a transformative purpose distinct from the original news articles. Plaintiffs report the news and seek to engage readers. Browse, in contrast, is an interactive research tool. In response to a user request, Browse finds and analyzes publicly available online content, extracts facts, and generates an original response with citations to relevant resources.

Making internal use of copyrighted content that does not allow users to access the copied material, but instead extracts and uses uncopyrightable elements of the work, is a recognized fair-use-favoring purpose. *See Romanova*, 138 F.4th at 115-18 (cataloguing fair-use-favoring justifications for copying). Numerous cases have held as much. *See, e.g.*, *Assessment Techs. of WI, LLC v. WIREdata*, 350 F.3d 640, 645 (7th Cir. 2003) (fair use to copy facts from a copyrighted work, "even if the [facts] were so entangled with the [work] that they could not be

extracted without making a copy of the [work]"); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520-28 (9th Cir. 1992) (fair use to make internal copy of copyrighted gaming software to extract uncopyrightable information even to create a competing video game); *Sony Comput. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 603, 606-07 (9th Cir. 2000) (fair use to make internal copy of software to "gain access to the unprotected functional elements" therein for purposes of creating non-infringing competitive product). In this regard, to the extent it provides factual information in response to user queries, Browse serves an even more transformative purpose than the search tool found to be highly transformative in *Google Books*, as it not only retrieves information behind the scenes but is capable of synthesizing its factual content for the user and providing pointers for further exploration. *See* 804 F.3d at 209.

Freedom to summarize, recast, and reprint facts first published by others is a principle central to copyright law and the news industry alike. *See Int'l News Serv. v. Associated Press*, 248 U.S. 215, 231, 234 (1918) ("[T]he news element—the information respecting current events contained in the literary production—is not the creation of the writer, but is a report of matters that ordinarily are publici juris; it is the history of the day."); *Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.*, 166 F.3d 65, 70 (2d Cir. 1999) ("[Defendant] had every right to republish the facts contained in [plaintiff's] articles."). News Plaintiffs themselves admit that "[redacted]" SUF ¶ 215. This Court has already rejected one attempt by News Plaintiffs to prevent OpenAI from providing users with facts reported in their articles. *See* Op. at 42, *New York Times v. Microsoft*, No. 1:23-cv-11195-SHS-OTW (S.D.N.Y 2025), Dkt. No. 514 (holding "when facts are reported in a different arrangement, with a different sentence structure and different phrasing, the secondary

work does not purloin protected expression, and no copyright infringement has ensued"). It should do so again here.

b) Factor 2: The Asserted Works are published and factual.

This factor favors fair use for Browse for the same reasons described above for pretraining. *See supra* Section IV(A); *see also Sega*, 977 F.2d at 1524–26 (second factor favors fair use because even though the defendant "copied protected expression," it did so to glean unprotectable information embedded therein).

c) Factor 3: The amount of any use was reasonable.

Browse uses no more of any work than is reasonably necessary to inform ChatGPT's response to the user and is not designed to share protected expression. *See Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 98 (2d Cir. 2014) (assessing whether copying was "reasonably necessary"); *Google Books*, 804 F.3d at 222 ("[W]hat matters . . . is not so much the amount and substantiality of the portion used in *making a copy*, but rather the amount and substantiality of *what is thereby made accessible* to a public for which it may serve as a competing substitute.") (emphasis in original).[13] When a webpage returned its content to ChatGPT, Browse identified a limited excerpt of "relevant information" to help inform its response. SUF ¶ 153. These excerpts are not displayed to the user. Rather, the factual information in the excerpts received from multiple webpages is synthesized and used to inform ChatGPT's response to the user. SUF ¶¶ 149, 154. The model limits how many words from any single source can appear in its

[13] The vast majority—over 96%—of the instances in which News Plaintiffs claim that Browse made an infringing copy of their work are instances where OpenAI merely received search results from Bing, containing only a short snippet of the underlying website, during the first step in the Browse process. SUF ¶ 163. Those were used to determine which search results were relevant, and OpenAI did not send a request for content to the associated website. SUF ¶ 164.

response. SUF ¶ 156. OpenAI used "[no] more of the copyrighted work than necessary" to achieve its transformative purpose of providing facts through ChatGPT. *See HathiTrust*, 755 F.3d at 98.

d) Factor 4: OpenAI's alleged use did not harm News Plaintiffs.

Google Books forecloses Plaintiffs' contention that they have suffered cognizable fourth-factor harm from Browse. 804 F.3d at 223–25. The principal question in that case was whether internal copies Google made of the complete text of copyrighted books was fair use. *Id.* at 207. In that context, the Second Circuit confronted the question whether fourth-factor harm could be found when showing a "snippet" of an author's work might satisfy the reader's interest and result in a lost sale. *Id.* at 223–25. The court's answer was a resounding "no." *Id.* at 224. It explained: "the type of loss of sale envisioned above will generally occur only in relation to interests that are not protected by the copyright." *Id.* The key consideration was that "[a] snippet's capacity to satisfy a searcher's need for access to a copyrighted book will at times be because the snippet conveys a historical fact that the searcher needs to ascertain"—and because such facts are not part of the protected expression in the copyrighted work, the defendant "would be entitled, without infringement of [the] copyright, to answer the [user's] query." *Id*. Put another way, because summarizing, recasting, and reprinting facts is not a copyright infringement, *see INS*, 248 U.S. at 234; *Nihon*, 166 F.3d at 70, it cannot be a cognizable harm under Factor 4. So even if News Plaintiffs were right that Browse "paraphrases" or "summarizes" facts from their articles, it would not matter. *See supra* Section II(B).[14]

[14] Nor would "widespread" use of facts available on the public Internet by services like Browse cause cognizable harm. *See Campbell*, 510 U.S. at 590 (fourth factor evaluates "whether unrestricted and widespread conduct of the sort engaged in by the defendant . . . would result in a substantially adverse impact on the potential market."). Moreover, as discussed above, any publishers who wished to prevent visits to their website via Browse could simply opt-out through the industry-standard robots.txt protocol or implement a hard block. *See supra* Section IV(B)(1).

News Plaintiffs' experts speculate that ChatGPT could harm News Plaintiffs by substituting for "traditional" search engines while providing lower "click through rates" to News Plaintiffs' websites than those "traditional" search engines. SUF ¶¶ 206, 247–48, 253, 258, 264, 271–72. As an initial matter, there is no evidence that they have suffered a decrease in web traffic due to ChatGPT. SUF ¶¶ 240, 243–70. News Plaintiffs' expert did not calculate click through rates on ChatGPT and admitted that she did not "[redacted]" SUF ¶¶ 272–73. In any event, this is not a cognizable theory of harm. The theory boils down to the contention that the alleged harm was caused by conduct that is fundamentally permissible as a matter of law. News Plaintiffs' own expert suggests that they might lose clicks not because ChatGPT displays their expressive content but because its [redacted] and leaves users [redacted] SUF ¶ 271. Under *Google Books* and the other precedent discussed above, that is not the kind of harm or substitution that counts in the fair use analysis.

3. OpenAI's alleged use of Asserted Works for Browse was impliedly licensed.

In any event, to the extent OpenAI made copies of the Asserted Works as part of the Browse process before News Plaintiffs opted out of that process via robots.txt or via hard block, those copies were impliedly licensed.

Implied license is a "complete defense to a claim of copyright infringement." *Plaid Takeover, LLC v. Owens*, No. 23-cv-3014, 2023 WL 6541300, at *7 (S.D.N.Y. Oct. 6, 2023). Such a license "may be inferred based on silence where the copyright holder knows of the use and encourages it." *Id.* In other words, "consent given in the form of mere permission or lack of objection is [] equivalent to a nonexclusive license." *Keane Dealer Servs., Inc. v. Harts*, 968 F.

Supp. 944, 947 (S.D.N.Y. 1997); *see also De Forest Radio Tel. & Tel. Co. v. United States*, 273 U.S. 236, 241 (1927); *Lukens Steel Co. v. Am. Locomotive Co.*, 197 F.2d 939, 941 (2d Cir. 1952). For decades, search engines like Google—which have built web indexes by crawling (*i.e.*, copying) virtually every webpage on the Internet—and other web-based products and services have relied on scraping or crawling the Internet. SUF ¶¶ 170–71. Indeed, News Plaintiffs use web scrapers of their own. SUF ¶ 172.

This copying has long been understood to be impliedly licensed unless the website owner opts out via mechanisms like the robots.txt protocol. *Field v. Google Inc.*, 412 F. Supp. 2d 1106, 1113–16 (D. Nev. 2006); *Parker v. Yahoo!, Inc.*, No. 07-cv-2757, 2008 WL 4410095, at *3–4 (E.D. Pa. Sept. 25, 2008).

There is no dispute that OpenAI publicly announced the Browse user agent and instructed the world on how to block it. *See* SUF ¶ 183. And there is no genuine dispute that News Plaintiffs were aware of OpenAI's announcement and chose not to act. *See* SUF ¶¶ 186–194. Under these circumstances, OpenAI was entitled to "infer that [News Plaintiffs] consent[ed] to [this] use," *De Forest Radio*, 273 U.S. at 241,[15] and therefore, to the extent Browse copied any Asserted Works before News Plaintiffs chose to update their robots.txt files or otherwise block Browse, those copies were impliedly licensed. *See id.; Field*, 412 F. Supp. 2d at 1116; *Parker*, 2008 WL 4410095, at *3–4.

[15] *Associated Press v. Meltwater U.S. Holdings, Inc.*, is not to the contrary. 931 F. Supp. 2d 537 (S.D.N.Y. 2013). Unlike OpenAI, the defendant there failed to "offer[] any expert testimony about robots.txt," which led the court to assume that robots.txt was an on-off switch that either permitted or prohibited all types of crawling. *See id.* at 563 (assuming that robots.txt makes "no distinction between" different kinds of uses). That is incorrect, particularly here, where OpenAI has set up an array of distinct user-agents so that publishers might do exactly what the *Meltwater* court suggested: "communicat[e] which *types* of use the copyright holder is permitting the web crawler to make of the content." *Id.* (emphasis in original); SUF ¶¶ 176, 183, 185.

C. News Plaintiffs' DMCA claims fail as a matter of law.

In addition to their copyright infringement claims, News Plaintiffs assert a claim under Section 1202(b)(1) of the Digital Millennium Copyright Act ("DMCA"). That section prohibits "intentionally remov[ing] or alter[ing] any copyright management information" ("CMI") with "reasonable grounds to know" that such removal would "induce, enable, facilitate, or conceal [copyright] infringement[.]" 17 U.S.C. § 1202(b). News Plaintiffs claim that OpenAI violated this section during the processing phase of LLM training, during which OpenAI extracts and converts raw data from a website into human language for pretraining. They are wrong.

1. Technical Background for DMCA claims

OpenAI trains its LLMs using data that comes from webpages on the public Internet. SUF ¶ 336. But the raw data looks nothing like what a user sees when she visits a website. SUF ¶ 336. Rather, it consists of thousands of lines of HTML code, graphics specifications, and other, non-language data that is illegible to most users. SUF ¶ 337. Data that News Plaintiffs allege might qualify as "CMI" is buried within and scattered throughout hundreds or thousands of fields of illegible data—*e.g.*, hidden metadata fields listing title or author; or fields containing HTML code that, when executed by a web browser, will display a "byline" at the top of a webpage. SUF ¶¶ 337–40.

Training a language model on illegible, non-language data would result in a model that primarily generates illegible, non-language data. SUF ¶ 341. Avoiding the use of "unnatural" language data for training is a well-established and necessary part of NLP ("Natural Language Processing") research, as confirmed by numerous technical papers (including those published by News Plaintiffs' own experts). SUF ¶ 342. So OpenAI ran computer scripts to identify and ***extract usable language data*** for model training. SUF ¶ 343. The purpose of OpenAI's text-extraction scripts was to convert raw HTML files into chunks of natural language text suitable

for model training—not to remove CMI. *See, e.g.*, SUF ¶ 344. Indeed, the relevant OpenAI source code for this process uses a series of heuristics to identify and ***retain*** the useful block of human language data on a page—not to identify and ***remove*** information elsewhere on the page. SUF ¶ 344–45. As a result, some alleged CMI for some works may not be extracted from the raw files since it is not useful for model training. SUF ¶ 346.

2. Legal Argument for DMCA claims

To establish a violation of Section 1202(b)(1), News Plaintiffs must establish (1) that they "conveyed" CMI "in connection with" their asserted works, (2) that OpenAI "remove[d]" or "alter[ed]" that CMI; (3) that "the removal was intentional;" and (4) that the defendant had "reasonable grounds to know[] that [the removal] will induce, enable, facilitate, or conceal an infringement of [copyright]." 17 U.S.C. § 1202(b)–(c); *Fischer v. Forrest*, 968 F.3d 216, 223 (2d Cir. 2020). Even assuming for purposes of this motion that News Plaintiffs conveyed CMI in connection with some of the Asserted Works, and that OpenAI's text-extraction scripts did not always extract all data that might qualify as CMI, News Plaintiffs' claim nevertheless fails for three independent reasons.

a) Any alleged CMI removal facilitated fair use, not infringement.

OpenAI's text-extraction process does not "induce, enable, facilitate, or conceal" copyright infringement. *See* 17 U.S.C. § 1202(b). Rather, it is an essential step of the pretraining process, which, for all the reasons described above, is fair use. Because fair use is "not an infringement of copyright," *Id.* § 107, any alleged CMI removal in service of a fair use cannot violate Section 1202. *See id.* § 1202(b). *Kadrey v. Meta Platforms, Inc.*, 2025 WL 1786418, at *1–2 (N.D. Cal. June 27, 2025) ("It does not make sense that Congress would have wanted to

exempt secondary users who make a fair use from [copyright] infringement liability, only to open them back up to DMCA liability if they removed some boilerplate in doing so.").

b) News Plaintiffs cannot establish "intentional removal" of CMI.

News Plaintiffs have failed to present evidence that would allow a reasonable fact-finder to conclude that OpenAI "intentionally remove[d]" CMI. 17 U.S.C. § 1202(b). This requires more than showing OpenAI ***knew*** its actions would result in the removal of CMI. This is clear from the statute itself, which uses *intent*-based language to define its first scienter requirement ("***intentionally*** remove or alter") while using ***knowledge***-based language to define its second scienter requirement ("having ***reasonable grounds to know***"). Indeed, Congress specifically rejected an earlier proposal for a knowledge-based standard. *Compare* NII Copyright Protection Act of 1995, H.R. 2441, 104th Cong. § 1202(b) ("No person shall . . . *knowingly* remove [CMI] . . .", *with* 17 U.S.C. § 1202(b) ("No person shall . . . *intentionally* remove [CMI] . . ."); *see also Zalewski v. Cicero Builder Dev., Inc.*, 754 F.3d 95, 107 (2d Cir. 2014); *Zuma Press, Inc. v. Getty Images (US), Inc.*, 349 F. Supp. 3d 369, 376 (S.D.N.Y. 2018).

News Plaintiffs' assertion that OpenAI utilized an extraction process that incidentally omitted CMI does not suffice as a matter of law. The Second Circuit's recent decision in *McGucken v. Shutterstock, Inc.*—concerning a Section 1202 claim against an image hosting platform that allegedly removed CMI from all uploaded images—is instructive. 166 F.4th 361, 370 (2d Cir. 2026). There, the court affirmed the grant of summary judgment in defendants' favor because the unrebutted evidence suggested that the purpose of the defendant's action was not to remove CMI specifically, but rather to "remove[] metadata from all images primarily to avoid computer viruses and the dissemination of personally identifiable information, and not for any reason related to infringement." *Id.*; *see also Beaulier v. Meta Platforms, Inc.*, No. 26-cv-

02632-CRB, 2026 WL 2521848, at *4 (N.D. Cal. Aug. 26, 2026) ("[R]etaining useful portions of information and discarding the rest is not sufficient to demonstrate intent."); *Logan v. Meta Platforms, Inc.*, 636 F. Supp. 3d 1052, 1064 (N.D. Cal. 2022) ("automatically omitting CMI" by separating copyrighted work from the "full context of the webpage where the CMI is found" is not sufficient to prove "intentionality as required by the DMCA").

Here, the undisputed evidence shows that the intended purpose of OpenAI's text-extraction scripts was to convert raw HTML files into chunks of natural language text suitable for model training—not to remove CMI. SUF ¶ 344. This is a standard and necessary step in NLP research. SUF ¶ 342. Even if a "side effect" of that process was to skip alleged CMI from some of the Asserted Works, that is not sufficient to create a triable issue of fact on News Plaintiffs' claim. *See McGucken*, 166 F.4th at 370; *Beaulier*, 2026 WL 2521848, at *4 (dismissing a claim based on "an instance of uniform transformation that incidentally sheds CMI"); *Logan*, 636 F. Supp. 3d at 1064.

c) News Plaintiffs cannot establish that any alleged CMI removal "concealed" infringement.

Even if News Plaintiffs could show that OpenAI "intentionally remove[d]" CMI, their claim still fails because they cannot show that OpenAI had "reasonable grounds to know" that any such removal would "induce, enable, facilitate, or conceal an infringement of [copyright][.]" 17 U.S.C. § 1202(b). News Plaintiffs have advanced two theories that purport to satisfy this requirement. Both fail as a matter of law.

First, News Plaintiffs argue that OpenAI's text extraction facilitated the pretraining process. *See* SUF ¶ 343. That theory fails because the pretraining process is a fair use, not infringement.

Second, News Plaintiffs argue that OpenAI's text extraction "concealed" infringement based on a hypothetical chain of inferences unsupported by the evidence. They speculate that, but for OpenAI's text-extraction process, ChatGPT would regurgitate CMI alongside regurgitations of Asserted Works, thereby informing users that News Plaintiffs own the Asserted Works. *See, e.g.*, The Times Third Am. Compl. ¶ 126, Dkt. No. 1677. From this, News Plaintiffs speculate that OpenAI would have had reasonable grounds to know that its text-extraction process would "conceal" that ChatGPT outputs are infringing copies of the Asserted Works. *Id.* ¶ 189.

News Plaintiffs' speculative theory fails at the outset because regurgitation hardly ever happens in the first place. But even in the exceptionally rare examples of alleged regurgitation they have identified, the source of the text is often apparent. The methods used by News Plaintiffs' experts to trigger instances of alleged regurgitation required them to already possess a copy of the Asserted Work (including CMI), portions of which (including title, author, or article text) were used in prompts to force OpenAI's LLMs to regurgitate a portion of that work. SUF ¶¶ 229–39. As a matter of logic, an output cannot conceal attribution information when the user draws upon that very information to prompt the output. To further underscore this common-sense point, OpenAI's expert analyzed all 68 alleged "regurgitations" identified in the parties' opening expert reports (across 108 million conversations) and made unrebutted findings that, in those rare instances where regurgitation occurs, the source of the allegedly regurgitated text would have been obvious to a user most of the time (63 of 68 or 92.7% of the identified instances). SUF ¶ 347. Thus, no reasonable juror could conclude that OpenAI had "reasonable grounds to know" that its text-extraction process would "conceal" alleged infringement in ChatGPT outputs.

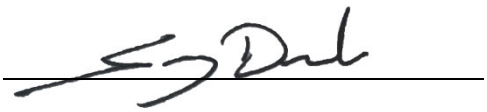
Nor is there evidentiary support for News Plaintiffs' speculation. The sole basis for News Plaintiffs' argument is their own expert reports that simply restate their speculation as a conclusion. These experts did not conduct a single experiment to support this speculation, nor do they cite a single third-party study suggesting that retaining CMI in training data would enable a model to generate accurate CMI alongside regurgitated body text–a fundamental premise of their theory. *See, e.g.,* SUF ¶ 349. Conclusory opinions by experts are not sufficient to defeat a motion for summary judgment. *Major League Baseball Properties, Inc. v. Salvino, Inc.*, 542 F.3d 290, 311 (2d Cir. 2008) (affirming summary judgment and rejecting conflicting expert testimony that was "neither accompanied by any evidentiary citation nor followed by any elaboration").

V. CONCLUSION

For the foregoing reasons, OpenAI respectfully requests that the Court grant this motion.

Dated: September 4, 2026

Respectfully submitted,



LATHAM & WATKINS LLP
Andrew M. Gass (*pro hac vice*)
andrew.gass@lw.com
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
Telephone: 415.391.0600

Sarang V. Damle
sy.damle@lw.com
Luke A. Budiardjo
luke.budiardjo@lw.com
1271 Avenue of the Americas
New York, NY 10020
Telephone: 212.906.1200

Elana Nightingale Dawson (*pro hac vice*)
elana.nightingaledawson@lw.com
555 Eleventh Street, NW, Suite 1000
Washington, D.C. 20004
Telephone: 202.637.2200

MORRISON & FOERSTER LLP
Joseph C. Gratz (*pro hac vice*)
jgratz@mofo.com
Tiffany Cheung (*pro hac vice*)
tcheung@mofo.com
Caitlin Sinclaire Blythe (*pro hac vice*)
cblythe@mofo.com
425 Market Street
San Francisco, CA 94105
Telephone: 415.268.7000

Lily Li (*pro hac vice*)
YLi@mofo.com
755 Page Mill Road
Palo Alto, CA 94304
Telephone: 650.813.5600

KEKER, VAN NEST & PETERS LLP
Robert A. Van Nest (*pro hac vice*)
rvannest@keker.com
R. James Slaughter (*pro hac vice*)
rslaughter@keker.com
Paven Malhotra
pmalhotra@keker.com
Michelle S. Ybarra (*pro hac vice*)
mybarra@keker.com
Reid P. Mullen (*pro hac vice*)
rmullen@keker.com
Nicholas S. Goldberg (*pro hac vice*)
ngoldberg@keker.com
633 Battery St.
San Francisco, CA 94111
Telephone: 415.391.5400

CERTIFICATE OF COMPLIANCE

In accordance with Local Civil Rule 7.1(c), I hereby certify that the foregoing OPENAI'S MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT OF MOTION FOR SUMMARY JUDGMENT is 8,185 words, exclusive of the caption page, table of contents, table of authorities, and signature block. The basis of my knowledge is the word count feature of the word-processing system used to prepare this memorandum.

Dated: September 4, 2026 By: