October 30, 2025

VIA ECF
Hon. Ona T. Wang
U.S. Magistrate Judge for the Southern District of New York
Daniel Patrick Moynihan United States Courthouse
500 Pearl Street
New York, NY 10007

RE: *In re OpenAI, Inc. Copyright Infringement Litig.*, No. 1:25-md-03143, This document relates to 23-cv-11195

Dear Judge Wang:

OpenAI submits this letter pursuant to the Court's order at ECF 715, which requested briefing on two issues: (1) the relationship of News Plaintiffs' motion docketed at ECF 656 and News Plaintiffs' spoliation allegations; and (2) the nature of the dispute underlying that motion and the status of the parties' agreements, if any, regarding the same.

Regarding the first request: the motion docketed at ECF 656 does not relate to News Plaintiffs' spoliation allegations or the broader issue of **data preservation**—which concerns whether OpenAI was required to deprive its users of the right to delete their own personal conversations. Instead, it relates to **conversation sampling**, *i.e.*, the question of how the parties will retrieve a sample of conversation data to satisfy News Plaintiffs' discovery requests, including (i) the size of the sample and (ii) most recently, whether OpenAI should be required to hand over tens of millions of irrelevant conversation logs. While both disputes arise out of the Times and its co-plaintiffs' repeated disregard for the privacy rights of ChatGPT users to gain tactical litigation advantages, the **data preservation** and **conversation sampling** disputes are distinct and should be treated as such.

Regarding the second request: as noted, the instant motion relates to the parties' efforts to derive a sample and search for potentially responsive documents from the conversation data that OpenAI stores in the ordinary course. The parties have already agreed on Phase 1 of this work, including (a) the date range for sampling (12/2022–11/2024), (b) the sample size (20 million logs), and (c) the methodology for retrieving these logs (*see* ECF 436-1). The impasse relates to Phase 2: how the parties will identify any relevant and responsive samples from the 20M.

News Plaintiffs demand that OpenAI hand over *the entire* 20M log sample "in readily searchable format" via a "hard drive or [] dedicated private cloud." ECF 656 at 3. That would include logs that are neither relevant nor responsive—indeed, News Plaintiffs concede that at least 99.99% of the logs are irrelevant to their claims. OpenAI has never agreed to such a process, which is wildly disproportionate to the needs of the case and exposes private user chats for no reasonable litigation purpose. In a display of striking hypocrisy, News Plaintiffs disregard those users' privacy interests while claiming that ***their own chat logs*** are immune from production because "it is possible" that their employees "entered sensitive information into their prompts." ECF 475 at 4. Unlike News Plaintiffs, OpenAI's users have no stake in this case and no opportunity to defend their information from disclosure. It makes no sense to order OpenAI to hand over millions of irrelevant and private

conversation logs belonging to those absent third parties while allowing News Plaintiffs to shield their own logs from disclosure.

To avoid that, OpenAI has proposed that the parties minimize the intrusion into users' privacy by using search terms and other standard discovery techniques (as the parties have been doing in the case) to identify a targeted set of ***relevant*** logs to produce under confidentiality protections. *See* Appendix A. This is fully consistent with the Court's instruction—and the Times's own statements—about the importance of protecting privacy interests. 5.27.2025 Tr. at 4:1–12.[1] OpenAI hopes to discuss this proposal at the upcoming October 31, 2025 settlement conference. In the meantime, the Court should deny News Plaintiffs' request.

1. Plaintiffs' Motion Relates to Conversation Sampling, Not Plaintiffs' Spoliation Allegations

The motion pending before this Court, *see* ECF 656 ("Motion"); ECF 679 ("Opp."), relates to ***conversation sampling***, which is distinct from ***data preservation***.

OpenAI's conversation data records are compressed, which means they are not searchable or viewable. ***Conversation sampling*** refers to the process for taking a sample of those records and decompressing them so that the sample of records can then be searched for relevant logs that are responsive to Plaintiffs' discovery requests (*e.g.*, evidence, if any, of "regurgitation" of Plaintiffs' copyrighted content). Previously, the parties have discussed (and briefed) how that sample should be created—a challenging technical issue. After over a year of negotiations, the parties agreed in August 2025 to move forward with creating a representative sample of 20M conversation logs. That sample will be ready in the next few weeks. The next step is to identify potentially responsive logs from this 20M sample. The parties' disagreement over this step is the dispute set out in the motion at ECF 656.

Data preservation relates to the allegation (raised during the discussions of conversation sampling) that OpenAI spoliated evidence by honoring users' preferences to delete their conversation logs after the Times filed suit in December 2023. OpenAI made clear it planned to do so in February 2024, *NYT* ECF 379-3, but News Plaintiffs ignored that disclosure and instead waited over 10 months to raise an objection via motion, *see NYT* ECF 379. This Court then issued a Preservation Order on May 13, 2025, after which OpenAI began preserving all ChatGPT prompt and output data and storing that data in segregated 30-day tables. The parties then ran hit counts on those segregated tables to determine whether OpenAI's non-retention of certain data caused a bias in what would be sampled. Consistent with the Court's instruction, this preserved data was not "produced to the plaintiffs or made public in any way, shape, or form." 5.27.2025 Tr. at 3:18–25. The parties completed that process and then stipulated to terminate the Order's obligations. ECF 628.[2] Again, none of the data in these segregated tables is part of the 20M.

---

[1] New York Times Editorial Board, *Privacy Cannot Be A Casualty of the Coronavirus* (Apr. 7, 2020) (arguing that people "should be able to control what happens to their personal data").

[2] News Plaintiffs suggest that OpenAI "had been deleting some historical ChatGPT output log data during the pendency of the Preservation Order." Mot. at 1–2. OpenAI did exactly what the Preservation Order directed: it created segregated tables of preserved data and set them aside for this litigation. ECF 33 (Court:

The dispute currently before the Court is therefore distinct from the parties' prior motion practice and discussions regarding data preservation: indeed, neither party has proposed further sampling or analysis of the preserved data. Instead, the Motion relates entirely to the plan for ***conversation sampling***.

2. OpenAI Did Not Renege on Any Agreement Regarding Conversation Sampling

Conversation sampling must proceed in two phases: **(1)** retrieving and decompressing a representative sample from OpenAI's compressed data storage, and **(2)** identifying the responsive and relevant documents from within that larger sample. The first phase (now almost complete) required the parties to negotiate and agree on an appropriate sample size and a process for retrieving the sample. The second phase (which the parties are currently addressing for the first time) will require the parties to determine how to identify the logs within that sample that are responsive to News Plaintiffs' requests and relevant to this litigation. This ensures that the logs produced to News Plaintiffs will be relevant and proportionate, as required by Rule 26.

Phase 1: Sampling. The parties' prior discussions on conversation sampling have related exclusively to the first "sampling" phase, focusing primarily on the parties' longstanding disagreement as to the appropriate sample size. A brief history of that dispute is below.

- The conversation sampling issue arose out of News Plaintiffs' May 2024 requests for inspection and production of "[q]uery, session, and chat logs ***related to [their] content***." *See, e.g.*, *NYT* ECF 379-8 (emphasis added); *NYT* ECF 288-1; *DN* ECF 135-1.
- After News Plaintiffs served those requests, OpenAI explained that, because its conversation data is stored in compressed format, the parties would need to derive a representative sample of conversation data to decompress so that it could be searched.
- To move that process forward, OpenAI supplied News Plaintiffs with a metadata list of all stored conversations and repeatedly requested that News Plaintiffs propose a methodology for sampling from that list that might satisfy their requests, indicating all along that the process would take significant time and effort.
- On May 20, 2025—one full year after News Plaintiffs served their initial Requests—News Plaintiffs finally proposed a sampling methodology. *See* ECF 394-3 at 3–4. That proposal demanded a sample of over 1.4 ***billion*** conversation logs. *Id.*
- OpenAI explained that the relevant statistical formula required a sample no larger than 5 million conversation logs, but offered 20 million as a compromise. ECF 394-4 at 1.
- News Plaintiffs then filed a motion asking OpenAI to retrieve 120 million logs, *see* ECF 394, before finally agreeing to proceed with a 20 million log sample, *see* Mot. at 2.
- On August 11, 2025, News Plaintiffs furnished OpenAI with a list of 19,999,992 conversations to retrieve, randomly sampled from the metadata list OpenAI produced.

"segregate and preserve"); 5.27.2025 Tr. at 29:9–17, 35:15–37:5 (OpenAI explaining preservation of "30-day safety tables"); ECF 55 (OpenAI confirming that the "mechanisms it has implemented" are "sufficient"). OpenAI did not interrupt its ordinary-course deletion processes because the Order did not require it to. 5.27.2025 Tr. at 34:17–20 (Court: "I'm not saying stop all your deletion processes.").

- OpenAI immediately began the process of retrieving, decompressing, processing, and de-identifying those 20M logs. *See* ECF 435-1 ¶3 (describing this process and estimating that it would "require up to 12 weeks" to process a 20 million log sample).
- OpenAI expects to complete that process in the next few weeks, as predicted. ECF 683.

Phase 2: Responsiveness. The above-described negotiations, which were the subject of prior motion practice, related entirely to the "sampling" phase. The Motion pending before the Court relates to the second phase, *i.e.*, how the parties will determine what subset of the 20M logs are actually responsive to Plaintiffs' requests. News Plaintiffs demand wholesale production of the entirety of the 20M-log sample. OpenAI disagrees and argues that the parties should first identify the potentially ***relevant*** and ***responsive*** conversations within the sample using standard search techniques the parties have already deployed in other areas of this case.[3]

3. Wholesale Production of 20M Logs Is Inappropriate

News Plaintiffs' demand for wholesale production of the 20M conversations is inappropriate. To be sure, if there is reason to believe that a conversation is relevant to this case and responsive to News Plaintiffs' requests—*e.g.*, because it contains a few sentences regurgitated from a *New York Times* article—OpenAI can make it available under the appropriate confidentiality protections (the "**Responsive Records**"). But there is no basis for News Plaintiffs to demand that OpenAI produce conversation records that are ***not*** relevant to any of the claims or defenses at issue here (the "**Excluded Records**"). Doing so would needlessly compromise the privacy interests of millions of users around the world while serving no benefit whatsoever to this litigation. That is not required by or consistent with Rule 26.

As a threshold matter, none of News Plaintiffs' discovery requests cover the Excluded Records. Those requests are all limited to "chat logs . . . related to [News Plaintiffs'] Content." *NYT* ECF 288-1; *DN* ECF 135-1. That excludes conversation data that is *not* related to their content. They concede that these unrelated logs constitute the vast majority (more than 99.99%) of the records contained within the 20M log sample. *See, e.g.*, ECF 394 at 2–3 (News Plaintiffs estimating that between 0.001% and 0.006% of ChatGPT logs are relevant).

More to the point: production of those Excluded Records would mean the disclosure of private conversations belonging to millions of ChatGPT users that have nothing to do with any issue in these cases. News Plaintiffs brush aside those privacy concerns by noting that the conversation logs will have already been processed via OpenAI's internal de-identification tool. Mot. at 2. But that tool targets categories of *identifying* information like names and addresses. *See* ECF 683 ¶3 It does not target *non-identifying* information that may nonetheless be private, like a user drafting an intimate letter to a partner, an individual's request for information about how her immigration status will be affected by current policies, a finance professional's use of ChatGPT to develop a

[3] News Plaintiffs claim that OpenAI already agreed to produce the 20M conversation logs in the sample and is now reneging on that agreement. That is false: OpenAI never agreed to do any such thing. News Plaintiffs cite OpenAI's June 25, 2025 letter, *see* ECF 394-4, but that letter discusses only how the sample would be created, not what would subsequently happen to the sample. They also cite OpenAI's stray use of the word "provide" in describing the sample size it could agree to generate in its July 18, 2025 letter, *see* ECF 394-6, but that falls far short of a commitment to hand over millions of user logs.

proprietary trading strategy, or a *Washington Post* reporter's use of ChatGPT to assist in the preparation of a news article. *Id.*; ECF 475 at 5 (NYT suggesting that such logs are protected by reporters' privilege); *Gonzales v. Google, Inc.*, 234 F.R.D. 674, 687 (N.D. Cal. 2006) (rejecting argument that "there are no privacy issues" absent "identifiable information").

Judge Parker's opinion in *Nichols v. Noom* is instructive. No. 20-cv-3677, 2021 WL 1997542 (S.D.N.Y. May 18, 2021). There, plaintiffs sought to compel production of chat logs from Noom's automated weight-loss coach for analysis. *Id.* at *1. The Court agreed with Noom's argument that plaintiffs' demand would require production of a "potentially large amount of highly personal, irrelevant information" about users' "weight-loss journey, food anxieties, life stressors," and so on. *Id.* at *1. To avoid that, the Court ordered Noom to use search terms on a random sample of 2,500 chats to identify the "subset" relevant to the plaintiffs' claims so that they could be produced. *Id.* at *3. But the Court held that the plaintiffs "are not entitled to" the "remainder of the 2,500" chatlogs—*i.e.*, those that did not hit on search terms—because "they do not contain relevant communications within the meaning of Rule 26." *Id.*

4. The Parties Should Collaborate To Identify Responsive Records

OpenAI remains committed to finding a similar solution here—*i.e.*, a solution by which it can respond to News Plaintiffs' discovery requests while minimizing the needless intrusion on its users' privacy interests. For example, Exhibit A contains a chart outlining the four categories of "different types of output [logs]" that News Plaintiffs claim they need to search for (as News Plaintiffs disclosed in their filings at ECF 394-1 and ECF 656-3), along with a specific proposal for how OpenAI could address each. OpenAI has confirmed that it has pre-existing tools that can be used to accomplish each of these proposed searches (or metadata analyses). ECF 683 ¶6. Prior to this filing, OpenAI sent the below proposal to News Plaintiffs with yet another request to meet-and-confer.[4] News Plaintiffs responded just one hour later, rejecting the proposal without offering any counter-proposal.

News Plaintiffs' Motion raised a vague objection that collaborating with OpenAI in this manner would require them to "reveal[] work product." Mot. at 3. But using search terms proposed by the requesting party is a well-established method for identifying potentially relevant documents that both parties have used extensively in this litigation. Indeed, the parties already engaged for months on a process through which, at the request of News Plaintiffs, OpenAI agreed to "run [] agreed N-gram search[es]" on its training datasets and "produce the results." *See NYT* ECF 462 at 2. The process OpenAI has proposed is no different. This is exactly "how discovery should work." Mot. at 1; *Nichols*, 2021 WL 1997542, at *2 ("Search terms are often used to cull emails and in fact are being used by the parties in this case [to do so].").

The parties should collaborate on a process to identify the Responsive Records so that OpenAI can make them available for this litigation, with the proper confidentiality protections. The Court should deny the Motion and instruct News Plaintiffs to confer with OpenAI in good faith regarding that process, including at the upcoming conference.

[4] The proposed search in the first row of the chart below—"news-related use cases"—does not fall within the contours of the discovery requests discussed above. But OpenAI is willing to work with News Plaintiffs to gather this information (which would not disclose the contents of private ChatGPT conversations) in light of the parties' ongoing discussions regarding classifiers.

Respectfully submitted,

KEKER, VAN NEST & PETERS LLP	LATHAM & WATKINS LLP	MORRISON & FOERSTER LLP
/s/ Edward A. Bayley	*/s/ Margaret Graham*	*/s/ Rose S. Lee*
Edward A. Bayley*	Margaret Graham	Rose S. Lee*

cc: All Counsel of Record (via ECF)

*All parties whose electronic signatures are included herein have consented to the filing of this document.