**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

| | |
|---|---|
| IN RE:<br><br>OPENAI, INC.,<br>COPYRIGHT INFRINGEMENT LITIGATION<br><br>This Document Relates To:<br><br>All Actions | 25-md-03143 (SHS) (OTW)<br><br>Hon. Sidney H. Stein |

**UPDATED MEMORANDUM OF LAW IN SUPPORT OF OPENAI'S RULE 72(a) OBJECTIONS TO THE ORDERS COMPELLING PRODUCTION OF CHATGPT CONVERSATION LOGS AT MDL ECF 734, MDL ECF 896, AND MDL ECF 910**

# TABLE OF CONTENTS

**Page**

# TABLE OF AUTHORITIES

**Page(s)**

## CASES



## RULES



## OTHER AUTHORITIES

## I. INTRODUCTION

This is a Rule 72 objection to a set of orders that required OpenAI to produce 20 million private ChatGPT user conversations to both the New York Times and its co-plaintiffs ("News Plaintiffs"), *see* ECF 734 (Initial Order); ECF 896 (Denying Reconsideration), and to the Class Plaintiffs, *see* ECF 910 § II(1)(iv) (collectively, the "Orders").[1] All parties agree that the 20 million conversations at issue include a massive amount of inherently private, sensitive, confidential, proprietary, and even privileged information belonging to ChatGPT users who have no role or stake in this case and no opportunity to object to the disclosure of this information in these proceedings. But as a result of those Orders, OpenAI will soon be forced to hand over this information to Plaintiffs and their experts, which would amount to one of the largest transfers of private personal information ever ordered by a United States District Court. The scale of this disclosure is irreconcilable with Rule 26's critical "proportionality" requirement. This Court should intervene immediately to stop it.[2]

The Orders suffer from two critical legal flaws. First, while the Court acknowledged that "the privacy considerations of OpenAI's users are sincere," the Court minimized those concerns

[1] OpenAI filed a Rule 72 objection to the order docketed at ECF 734 on November 24, 2025, ECF 841, which directed OpenAI to produce to the News Plaintiffs the 20 Million ChatGPT Logs at issue, and simultaneously moved the court below to reconsider that order, ECF 742. After the Court below declined OpenAI's motion for reconsideration, ECF 896, this Court ordered OpenAI to "file [this] updated memorandum in support of their objection" addressing the order denying reconsideration and the timeliness of OpenAI's original objection, ECF 903. In the interest of efficiency, this memorandum also addresses OpenAI's separately noticed Rule 72 objection to the Court's recent cross-production order in the Class case, ordering OpenAI to produce the 20 Million ChatGPT logs to the Class Plaintiffs as well. ECF 910 at 4.

[2] On December 8, 2025, OpenAI filed a letter requesting that Judge Wang hold the Orders in abeyance pending this Court's review of OpenAI's Rule 72 objections. ECF 909. Today, Judge Wang denied that request. ECF 932. Accordingly, absent prompt intervention from this Court, OpenAI will have no choice but to prepare to make the data available as soon as it is de-identified, likely in the coming days.

rather than balancing them as Rule 26 requires. ECF 896 at 7. It did so based on the assumption that the process OpenAI is currently using to "de-identify" the conversations is sufficient to eliminate any privacy risk. *Id.* at 7-8. That was factually wrong. As OpenAI has explained in two (unrebutted) sworn witness declarations, the de-identification process is not designed to remove information that is private but non-identifying — *i.e.*, the exact kind of information that the Court below held was immune from disclosure when denying OpenAI's motion to compel the Times to produce its own chatlogs. ECF 604. Put simply, just removing names, contact information, and numeric identifiers from a record in which a user also discusses details of their finances, immigration status, trading strategies, or investigative journalism does not mean that the record is no longer private, sensitive, confidential, or privileged. None of the Court's Orders explain why the disclosure of this type of private information should be immune from discovery when it was generated by the Times in its own chatlogs using OpenAI's models, but not when it was generated by innocent non-parties with no role, stake, or voice in these proceedings.

Second, the Orders failed to acknowledge OpenAI's far less burdensome proposal for satisfying the needs of this case without unnecessary disclosure of a massive trove of private non-party information. Over a month ago, OpenAI advanced a detailed proposal under which the parties would use techniques that they have used for months in other areas of this case to accomplish each and every one of Plaintiffs' user-data related discovery goals. ECF 717-1. The parties, for example, could use search terms and n-gram matching to identify and produce ChatGPT conversation logs that have some particularized relevance to Plaintiffs' claims or the works they have asserted here — a subset of logs that News Plaintiffs and their expert have

estimated comprise between 0.001 to 0.006% of ChatGPT conversations. ECF 394 at 2–3.[3] For the remaining 99.994–99.999% of logs that do not have any particularized relevance to the claims at issue here — and which are at best only generally relevant in the sense they show how ChatGPT is used for non-infringing purposes — the parties can conduct whatever analyses they need through non-content metadata like classifiers, which would not require the unnecessary disclosure of the private information in those conversation logs. *See* ECF 717 (Letter), 717-1 (Proposal). That proposal balances the needs of the case with the privacy interests of OpenAI's users.

None of the Orders even acknowledge that proposal, and News Plaintiffs have never contended that the proposal is unworkable or insufficient to meet the needs of this case. By ignoring that more reasonable alternative and instead ordering the production of all 20 million records notwithstanding their undisputed lack of specific, particularized relevance to the claims and defenses in this action, the Court below committed clear error. This Court should intervene to safeguard the privacy of OpenAI's users.

## II. BACKGROUND

### A. ChatGPT Conversation Data Discovery Requests

In May 2024, News Plaintiffs requested inspection and production of "[q]uery, session, and chat logs related to [their] Content," including "user queries paired with responses to those queries." ECF 864 Ex. A, at 2 (emphasis added); ECF 864 Ex. B, at 13; *see also* ECF 864 Ex. C at 2; ECF 864 Ex. D, at 17. OpenAI responded that the data was not readily searchable given the

[3] The Times's ChatExplorer logs—which are the subject of another Rule 72 objection pending before this Court, *see* ECF 639—have particularized relevance to this case for an additional reason: unlike the non-party logs that Plaintiffs seek here, the Times's logs are party admissions that rebut the Times's contention that OpenAI's models harm journalism and are not transformative.

format in which it was stored,[4] and so the parties would need to derive a representative sample that could be searched. *See* ECF 864 Ex. E. The parties continued to meet-and-confer thereafter for the next several months. *See* ECF 864 Exs. F–K (listing correspondence).

In June 2024, Class Plaintiffs similarly requested production of "the output of any Large Language Models created by [OpenAI] related to Plaintiffs' works." Supplemental Declaration of Edward Bayley[5] Ex. W (RFP No. 52) (emphasis added). OpenAI also met and conferred with Class Plaintiffs for several months, explaining that their proposal to "search the output data for the named plaintiff's works and/or their names" posed a question of "technical feasibility" because the data is not readily searchable. Supp. Bayley Decl. Ex. X at 17, 59.

**B. ChatGPT Conversation Data Sampling Dispute**

On May 20, 2025, News Plaintiffs proposed a sampling methodology contemplating a sample of more than 1.4 billion ChatGPT conversation logs. *See* ECF 864 Ex. L, at 3–4. Pursuant to meet-and-confers, in June 2025, OpenAI offered as a compromise to create a sample of up to

[4] Due to the significant volume of data generated by the hundreds of millions of ChatGPT conversations each month, OpenAI compresses conversation log data before placing it into long-term, "offline" storage. *See* ECF 43-18. After a conversation log is compressed, only a limited amount of metadata about the conversation, such as the date it was created, remains visible and in a searchable format. *See* ECF 864 Ex. K, at 2. To view and search the full content of the logs, the logs must first be decompressed. *Id*. In other words, OpenAI cannot, for example, search for every conversation in which a user mentioned "The New York Times" without first decompressing and processing every single one of the billions of ChatGPT conversations in offline storage. Because of these technical challenges, the parties have agreed that the appropriate way to test News Plaintiffs' output-based infringement claims is to decompress a sample of ChatGPT conversations so that statistically significant conclusions can be drawn about the full set of historical conversation data. *See, e.g.,* ECF 864 Ex. H.

[5] To avoid an additional sealing motion, this brief cites to the exhibits and attorney declaration that OpenAI previously filed with its initial Rule 72 objection (*see* ECF 864) rather than re-filing each exhibit. *See* ECF 903 (Order directing OpenAI to "file an updated memorandum of law"). OpenAI has also simultaneously filed a supplemental declaration with three additional exhibits to address the developments since the initial Rule 72 objection, including this Court's request that OpenAI address the timeliness of its initial filing under Rule 72. *See* Supplemental Decl. of Ed. Bayley.

20 million conversations from which responsive content could be searched. ECF 864 Ex. M. Although News Plaintiffs initially filed a motion seeking to compel OpenAI to retrieve a sample of 120 million conversations, *see* ECF 394, they eventually agreed to proceed with a 20 million sample. ECF 719-2.

In August, News Plaintiffs furnished OpenAI with a list of nearly 20 million conversations, selected from a metadata list OpenAI provided months earlier, to use as the sample for the searching of records related to News Plaintiffs' content. *Id*. OpenAI immediately began the extensive, complex and lengthy process of retrieving, decompressing, processing, and de-identifying each of those 20 million conversations (which in turn constitute multiple prompt-output pairs). *See* ECF 435-1 ¶3 (describing the decompression and de-identification process).

### C. Production of ChatGPT Conversations and the Challenged Orders

On October 14, 2025, News Plaintiffs demanded for the first time that OpenAI "commit to producing output log data" in its entirety. ECF 864 Ex. R, at 9. As of that date, the parties had not yet discussed which conversations within the sample would be made available in response to News Plaintiffs' discovery requests, or the logistics of making these conversations available. ECF 656 at 1. OpenAI explained to News Plaintiffs that—while it could not turn over 20 million conversations wholesale given user privacy concerns and the fact that the overwhelming majority are not responsive to News Plaintiffs' discovery requests—it would work with News Plaintiffs "to identify the [conversations] that are actually relevant to [their] claims" of output-based copyright infringement and make them available. ECF 864 Ex. R, at 8.

The next day, without responding to OpenAI's offer to meet and confer, News Plaintiffs filed a motion seeking to compel OpenAI to produce the entire 20 million sample. ECF 656 at 1. News Plaintiffs falsely asserted that OpenAI "reneged" on an agreement to produce the entire

sample, even though the parties had yet to discuss these specific issues, much less come to an agreement. News Plaintiffs have not, either in their initial motion or in subsequent briefing, explained why OpenAI's proposal was not sufficient to find the logs responsive to their requests.

At the October 29, 2025 discovery conference, Judge Wang instructed the parties to file simultaneous supplemental briefs explaining "the nature of the dispute" and the parties "agreements" regarding production of this data. ECF 715. After that conference, OpenAI attempted to resolve the dispute by proposing methods to News Plaintiffs to search for conversations related to their works and offered to discuss the proposal with News Plaintiffs in an attempt to reach a resolution. ECF 864 Ex. R, at 1–2. News Plaintiffs refused to engage with that proposal. *Id.* OpenAI then submitted a supplemental brief detailing the history of the dispute as Judge Wang requested, *see* ECF 717 at 2–4, and reminding the Court of the serious privacy concerns at stake. *See id.* at 4–5 (emphasizing the privacy implications of producing de-identified conversations).

But on November 10, 2025, the Court issued the Initial Order,[6] which, without discussion of responsiveness, relevance, or proportionality under Rule 26(b), directed OpenAI "to produce the 20 million de-identified Consumer ChatGPT Logs[.]" ECF 734 at 2. The Initial Order faulted OpenAI for failing to explain "how its consumers' privacy rights are not adequately protected" or why the district court's order in *Concord Music Grp., Inc. v. Anthropic PBC* requiring production of an entire 5-million record sample "is not similarly instructive here." *Id.* at 1–2; *see also Concord*

[6] Although the docket lists the Initial Order as being dated November 7, 2025, the Initial Order was not entered into the ECF system and thereby served on the parties until November 10, 2025. *See* Supp. Bayley Decl. Ex. Y ("Notice of Electronic Filing" stating that the Initial Order "was entered on 11/10/2025 at 9:28 AM EST and filed on 11/7/2025[.]").

*Music Grp. v. Anthropic PBC, No*. 24-CV-3811 (EKL) (SVK), 2025 WL 1482734, at *3–4 (N.D. Cal. May 23, 2025).

OpenAI promptly moved for reconsideration, explaining that it had not addressed the *Concord* order because News Plaintiffs first raised *Concord* in their simultaneous supplemental briefing—not in their original motion to compel at ECF 656—and therefore OpenAI had no opportunity to respond. ECF 742 at 2. OpenAI further explained that the order in *Concord* was inapposite because defendant-Anthropic *voluntarily* offered to produce the entire sample of queries and responses negotiated by the parties in that litigation. *Id.* at 3. As a result, the court in *Concord* did not consider the privacy considerations implicated by compelling production of millions of entire conversations as is the case here. *Id.*

On December 3, 2025, the Court denied OpenAI's motion for reconsideration. ECF 896. The Court found that all 20 million conversations are relevant to News Plaintiffs' output claims "to the extent that they contain partial or whole reproductions of [their] copyrighted works" and to OpenAI's defenses "to the extent they contain other user activity." *Id.* at 7. The Court also addressed proportionality for the first time, finding that production of the conversations would have "minimal burden" to OpenAI and its users despite recognizing the importance of the privacy interests at stake. *Id.* at 7–8.

At the December 4, 2025 discovery conference, the Court further directed OpenAI to cross produce all 20 million conversations to Class Plaintiffs[7] but stayed the order to permit OpenAI to file this objection. *See also* December 4, 2025 Hr'g Tr. at 113:11–18. OpenAI requested that the

[7] Class Plaintiffs did not separately move to compel production of ChatGPT conversations responsive to their requests or join News Plaintiffs' motion to compel at ECF 656. And Class Plaintiffs clarified in a submission on November 24, 2025 that they are "seeking *only the exact same material* which will be made available to News Plaintiffs following the final resolution of the dispute regarding OpenAI's production of output logs." ECF 869 at 5 (emphasis in original).

Court stay production to all Plaintiffs pending resolution of this Rule 72 objection, which the Court denied. *See* ECF 909, 932.

## III. STANDARD OF REVIEW

A district court may "modify or set aside" any part of a magistrate judge's non-dispositive order "that is clearly erroneous or is contrary to law." Fed. R. Civ. P. 72(a). An order is "contrary to law" if it "fails to apply or misapplies relevant statutes, case law, or rules of procedure." *Coventry Cap. U.S. LLC v. EEA Life Settlements Inc.*, 333 F.R.D. 60, 64 (S.D.N.Y. 2019) (quoting *R.F.M.A.S., Inc. v. So*, 748 F. Supp. 2d 244, 248 (S.D.N.Y. 2010)). An order is "clearly erroneous" if it provides the reviewing court a "definite and firm conviction that a mistake has been committed." *Id.*

## IV. ARGUMENT

The order below was clearly erroneous, as set forth below. *See infra* Section IV.A. Moreover, OpenAI's Rule 72 objection was timely under Rule 72(a) because OpenAI filed it exactly 14 days after being served with a copy of Judge Wang's order via the court's ECF system. *See infra* Section IV.B.

### A. The Orders Below are Clearly Erroneous

The Order below should be overturned for two independent reasons. First, the Order failed to adequately consider the privacy interests of absent non-parties. *See infra* Section IV.A.1. Second, the Order failed to consider (and did not even acknowledge) a far less burdensome means to satisfy the needs of this case. *See infra* Section IV.A.2.

#### 1. The Orders Failed To Afford Due Weight to the Privacy Interests of OpenAI's Non-Party Users

Every day, hundreds of millions of individuals around the world turn to ChatGPT for highly personal or sensitive purposes, such as drafting an intimate letter to a partner, requesting

information on one's immigration status, managing personal finances, seeking to develop a proprietary trading strategy, or conducting investigative journalism. As OpenAI has explained at length (and as Plaintiffs have never disputed), the ChatGPT conversation logs at issue here therefore contain highly sensitive private information — exactly the same kind of information that the Times itself has sought to protect from disclosure in this case. *See, e.g.*, ECF 475 at 5 (the Times asserting reporters' privilege over ChatExplorer logs). But unlike the Times's chatlogs, *see id.*, the ChatGPT conversations at issue here belong to non-parties who have no role or stake in these proceedings and no opportunity to object to the disclosure of information that they might consider private, confidential, proprietary, or even privileged.

The Second Circuit has instructed courts in this District to take these concerns seriously. In *S.E.C. v. Rajaratnam*, 622 F.3d 159 (2d Cir. 2010), the defendants appealed a district court discovery order compelling them to disclose thousands of wiretapped conversations. *Id.* at 164. In considering that appeal, the Second Circuit emphasized that "[t]he privacy interests in the instant case merit particular attention given that the disclosure order implicated thousands of conversations of hundreds of innocent parties." *Id.* at 184. The Court observed that "ordering the disclosure of *all* the conversations without limiting discovery to relevant material could infringe the privacy rights of hundreds of individuals, whose irrelevant, and potentially highly personal, conversations . . . would needlessly be disclosed to the SEC and other parties, without furthering any legitimate countervailing interest." *Id.* at 187 (emphasis in original). And on that basis, the Court held that the "district court clearly exceeded its discretion" by failing to "balance" the SEC's interests in disclosure "against the privacy interests at stake," which (according to the panel) would have required limiting the disclosure "to relevant conversations." *Id.*; *see also Nichols v. Noom*, No. 20-cv-3677, 2021 WL 1997542, at *3 (S.D.N.Y. May 18, 2021) (rejecting plaintiffs' demand

for an unfiltered production of chat logs to avoid the disclosure of a "potentially large amount of highly personal, irrelevant information" about users' "weight-loss journey, food anxieties, life stressors," and so on, and instead ordering narrowing via search terms).[8]

The Court below made the same error by ordering the "wholesale" production of 20 million conversations records. In the Orders under review, the Court minimized the privacy interests at stake rather than balancing them against Plaintiffs' need for disclosure. It did so by relying on the fact that OpenAI has been exercising best efforts to de-identify the conversation logs in advance of making them available. *See* ECF 896 at 8. But as OpenAI has explained in two sworn and unrebutted witness declarations, OpenAI's de-identification process is not designed to scrub information that is private but not directly identifying — like confidential source code, proprietary trading strategies, medical details, or financial information. *See* ECF 435-1 ⁋ 8, ECF 683 ⁋ 3.[9]

---

[8] The Court below held that *Noom* was inapposite, reasoning that unlike in *Noom*, all of the logs in the 20 million-log sample of ChatGPT consumer conversations are relevant to (e.g.) OpenAI's fair use defense in the sense that they are illustrative of broad patterns of ChatGPT usage. ECF 896 at 8. But while OpenAI's defenses do rely, at least in part, on the fact that ChatGPT is predominantly used for purposes that have no relation to Plaintiffs' works, that alone cannot be sufficient to justify the wholesale production of the private contents of ChatGPT conversations created by non-parties with no opportunity to object to such disclosure on privilege or privacy grounds. That is particularly so given that OpenAI has proposed a less burdensome and intrusive mechanism to investigate (e.g.) the breadth of non-infringing ChatGPT usage through classifiers and other similar means. *See infra* Section A.2.

Moreover, the Court below emphasized that Judge Parker's opinion in *Noom* held that the "Protective Order in place" was sufficient to assuage any privacy concerns arising from the plaintiff's review of the narrowed set of chatlogs that the defendant would be required to produce after narrowing for relevance. ECF 896 at 8-9 (citing *Noom*, 2021 WL 1997542, at *4). But that overlooks the fact that the dataset in *Noom* was limited to chats of users in the putative class. *See id.* at *2. Here, by contrast, the private information in the logs at issue are owned predominantly, if not entirely, by non-party ChatGPT users with no stake in this case.

[9] Indeed, it has been well-documented that de-identification is insufficient on its own to guarantee anonymity, as it may still be possible to trace de-identified information back to the individuals. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010); *see also* ECF 864 Exs. S–T. For instance, in August 2006, after America Online (AOL) publicly released the search queries for 650,000

This is exactly the type of private information that the Times itself cited when resisting, successfully (so far), production of its own logs. ECF 475 at 4–5.

Neither the Court's Initial Order nor the Order Denying Reconsideration explained why such non-identifying but private information should be immune from discovery when that information was generated by the Times, *see* ECF 604 (denying Defendants' motions to compel), but not when it was generated by innocent third parties with no role, stake, or voice in these proceedings. That gets it backwards: as the Second Circuit has instructed: "[t]he job of protecting [the privacy] interests [of absent non-parties] rests heavily upon the shoulders of the trial judge, since all the parties who may be harmed by disclosure are typically not before the court." *Matter of New York Times*, 838 F.2d 110, 116 (2d Cir. 1987).

The Court's Order denying reconsideration raised two additional grounds for minimizing the privacy interests of OpenAI's users, neither of which are availing. First, the Court noted that OpenAI has already "reduc[ed] the total output logs subject to discovery from tens of billions to 20 million." ECF 896 at 8. But that is cold comfort to the millions of users whose private information is laid bare in those 20 million ChatGPT conversation logs. The fact that disclosure of billions of private ChatGPT conversations would constitute an even more egregious privacy violation does not mean that the disclosure of tens of millions of conversations does not raise exigent privacy concerns. Second, the Court noted that the conversation logs may be protected by the Protective Order in this case. ECF 896 at 8. But as OpenAI has already explained, the existence of a protective order has never been a sufficient basis to brush off concerns over the

anonymized users of AOL's search engine, two of the Times' own reporters were able to deduce the identity of one of these individuals based on her de-identified search queries. ECF 864 Ex. T. As these reports illustrate, although de-identification is an important and necessary tool, it cannot fully address all of the substantial privacy concerns associated with producing millions of personal conversations of individuals who are not parties to this case. *See* ECF 683 at ⁋ 3.

disclosure of private information, especially of non-parties. ECF 752 at 2. The Second Circuit said as much in the *Rajaratnam* case: "[B]ecause it would not stop the initial disclosure to the SEC and the other parties, the district court's protective order limiting disclosure to the SEC and other parties to the litigation was not sufficient to protect the privacy interests at stake." 622 F.3d at 186 n.27.[10]

The Court's failure to adequately weigh the privacy interests of OpenAI's users was clear error. *See Rajaratnam*, 622 F.3d at 187-88. For that reason alone, OpenAI's objection should be sustained.

**2. The Orders Overlooked A Far Less Burdensome and Intrusive Alternative**

Worse still, the privacy burdens imposed by the Order are entirely unnecessary. Over one month ago, OpenAI proposed a far less burdensome mechanism through which OpenAI would use a suite of techniques—including traditional search terms, n-gram matching, and conversation classifiers—to identify relevant conversation logs "to produce under confidentiality protections." ECF 717. Below is a table memorializing this proposal, which OpenAI filed with the Court on October 30, 2025, as ECF 717-1. The column on the left quotes directly from a paragraph of News Plaintiffs' expert declaration submitted during the briefing on this issue, which lays out the "multiple different types of output in OpenAI's ChatGPT output logs" that "News Plaintiffs want to analyze." ECF 656-3 ¶ 3; *see also* ECF 875 at 4 (noting that OpenAI offered to "conduct n-gram searches and targeted keyword searches" for Class Plaintiffs and to "confer with [Class]

[10] The Court's reconsideration order also questioned why OpenAI would de-identify "the entire 20 million log sample" if "OpenAI never intended to produce the entire 20 Million ChatGPT Logs." ECF 896 at 9 n.9. OpenAI's regular practice, however, is to always de-identify its customers' conversations logs *before any* content-based analysis (e.g., n-gram searching) occurs. *See, e.g.*, Aaron Chatterji et al., *How People Use ChatGPT*, OpenAI (Sept. 15, 2025) at 5, 7, 9 (explaining that OpenAI de-identified 1 million ChatGPT messages for research, even though no member of the research team was permitted to set eyes on the contents of those messages).

Plaintiffs on search terms," and that Class Plaintiffs refused this offer without explanation). The column on the right proposes a methodology for satisfying each of News Plaintiffs' needs, as enunciated in that declaration.

| **News Plaintiffs' Proposed Searches** (ECF 656-3 ¶ 3) | **OpenAI's Proposed Methodology** |
|---|---|
| "the prevalence of prompts and outputs reflecting news-related use cases" | Use classifiers to identify prevalence of news-related use cases |
| "the prevalence of regurgitated outputs containing News Plaintiffs' copyrighted content" | Using n-gram matching to identify outputs containing text from News Plaintiffs' articles |
| "reproduction of News Plaintiffs' copyrighted content in the Retrieval Augmented Generation ('RAG') process" | Search specific RAG-related data fields for News Plaintiffs' domains |
| "the prevalence of hallucinated outputs diluting News Plaintiffs' famous trademarks" | Use targeted keyword searches to find instances of ChatGPT outputs that attribute facts to News Plaintiffs |

The motivation behind OpenAI's proposal is simple. The parties would use search terms and n-gram matching to identify ChatGPT conversation logs that have some particularized, specific relevance to the claims or works at issue in this case — *e.g.*, ChatGPT conversations that allegedly include some snippet of text from a *New York Times* article. Because those logs have specific relevance to Plaintiffs' claims, OpenAI would produce them under an appropriate confidentiality designation. But for the remainder of logs—which News Plaintiffs and their expert estimated would constitute between 99.994% and 99.999% of the sample, *see* ECF 394 at 2–3 (estimating prevalence rates of $p = 0.00001$ and $p = 0.00006$)—production is unnecessary. These logs do not relate to the substance of Plaintiffs' works, nor do they have particularized relevance to any of Plaintiffs' allegations that their works are being infringed. At most, they are generally relevant to OpenAI's defenses, including (as the Court below correctly noted) to the many transformative uses of OpenAI's models and the vast array of non-infringing uses of those models.

ECF 896 at 6-7.[11] But investigating the many ways in which OpenAI's models are used by users around the world does not require wholesale production of tens of millions of user logs. To the contrary, the parties can accomplish that goal while respecting the privacy concerns of non-parties through non-content metadata, like conversation classifiers (as the Court below has recognized in other contexts), *see* Tr. of 2025.05.27 Hr'g at 51:14-24 (directing the parties to "use [] classifiers and non-content information" to investigate OpenAI's API logs to avoid privacy concerns). That is precisely what OpenAI proposed. ECF 717-1.

None of these techniques are novel to this case: both News Plaintiffs and Class Plaintiffs have already requested and received classifier data regarding ChatGPT behavior; the parties have relied heavily on n-gram searching in the context of discovery into OpenAI's training data; and the parties have used search terms extensively for ESI discovery since the beginning of this case. Hon. Ona T. Wang, *Model Joint Electronic Discovery Submission and Proposed Order* (Mar. 8, 2018) § 6 ("The parties have discussed methodologies or protocols for the search and review of electronically stored information, as well as the disclosure of techniques to be used. Some of the approaches that may be considered include: the use and exchange of keyword search lists, 'hit reports,' and/or responsiveness rates[.]").

Plaintiffs have refused to even acknowledge this proposal in the five weeks since OpenAI advanced it. At no point has any plaintiff explained why the proposal would be unworkable or

[11] "[T]he fact that [these logs are] relevant does not mean that its production will always be proportional to the needs of the case." *Energy Transportation Grp., Inc. v. Borealis Mar. Ltd.*, 2024 WL 1436133, at *3 (S.D.N.Y. Mar. 22, 2024), *reconsideration denied*, 2024 WL 3318113 (S.D.N.Y. June 18, 2024). To the contrary, courts in this District consider the "marginal utility" of requested discovery when resolving such disputes, which often means denying requests for relevant information when their minimal or incidental relevance is outweighed by countervailing interests, like burden. *See Alaska Elec. Pension Fund v. Bank of Am. Corp.*, No. 14-cv-7126, 2016 WL 6779901, at *3–4 (S.D.N.Y. Nov. 16, 2016) (collecting cases).

insufficient to meet the needs of this case. News Plaintiffs' only response to this proposal has been to vaguely gesture to "work product" concerns. ECF 656 at 3. But as OpenAI explained, the proposed methods have already been used extensively in this litigation without either party raising work-product concerns. ECF 717 at 5. Since OpenAI made that point, News Plaintiffs have failed to raise work-product concerns when briefing this issue. *See, e.g.*, ECF 746.[12]

Put another way, it is uncontested that OpenAI's alternative is a less burdensome means to satisfy the needs of this case. Under OpenAI's proposed methodology, the parties would agree on search terms to identify and make available any conversation logs from within the 20 million-log sample that were "related to [plaintiffs'] content" — thereby fully satisfying the scope of their discovery requests. *See supra* Section II.A. (quoting Plaintiffs' discovery requests). And the parties could use classifiers to investigate and draw conclusions about the nature of ChatGPT usage more broadly, including the extent to which people use ChatGPT for "news-related use cases," *see* ECF 656-3 ¶ 3, or the extent to which people use ChatGPT for transformative and socially beneficial purposes that have nothing whatsoever to do with Plaintiffs' books or news articles.

The Court below inexplicably overlooked this proposal, both in its Initial Order (ECF 734) and its Order Denying Reconsideration (ECF 896), even though OpenAI has repeatedly raised this alternative in briefing this issue. *See, e.g.*, ECF 717; ECF 742. As a result, the Court failed to acknowledge or dispute that OpenAI's proposal would have fully satisfied the parties' needs for discovery into both (1) conversation logs "where News Plaintiffs' copyrighted works may be

[12] Courts in this district have likewise endorsed search strategies to identify responsive documents and have reprimanded plaintiffs' "recalcitrance" for refusing to agree to a search methodology to define the scope of production. *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 374 (S.D.N.Y. 2006); *see also Nichols*, 2021 WL 1997542, at *3 (ordering that "search terms be applied to identify relevant chats" and holding that "Plaintiffs do not actually need to see the chats for the remainder of the 2,500 users. . . . Indeed, they are not entitled to them, as they do not contain relevant communications within the meaning of Rule 26.").

reproduced in whole or in part," *see* ECF 896 at 6, which OpenAI's methodology would identify via n-gram matching, *see* ECF 717-1, and (2) conversation logs that are relevant to "damages and factor four of its fair use defense," *see* ECF 896 at 6, which OpenAI's methodology would quantify through classifiers — while minimizing the privacy burdens on non-parties. That, too, is clear error. Rule 26 is clear: "[o]n motion or on its own, the court must limit the frequency or extent of discovery" where the discovery sought can be obtained through "less burdensome" means. Fed. R. Civ. P. 26(b)(2)(C)(i). The Orders below should have considered OpenAI's "less burdensome" mechanism for dealing with this discovery.

### B. OpenAI's Objection Was Timely

This Court also instructed OpenAI to explain why its initial objection was timely filed. ECF 903. While the order below was dated November 7, 2025 (a Friday), OpenAI did not receive a copy of that order via the court's ECF system until November 10, 2025 (the following Monday). *See* Supp. Bayley Decl. Ex. Y (Notice of Electronic Filing). OpenAI filed its Rule 72 objection 14 days later, on November 24, 2025. ECF 841. That is fully consistent with Rule 72(a). *See* Fed. R. Civ. P. 72(a) (requiring the filing of objections "within 14 days after being *served* with a copy" (emphasis added)); *see also Thai Lao Lignite (Thailand) Co. v. Gov't of Lao People's Democratic Republic*, 924 F. Supp. 2d 508, 517 (S.D.N.Y. 2013) (referring to the "ECF filing notice" to determine the timeline for filing a Rule 72 objection).

## V. CONCLUSION

For the foregoing reasons, this Court should set aside the orders docketed at ECF 734, ECF 896, and ECF 910 § II(1)(iv).

Respectfully submitted,

*/s/ Margaret Graham*
**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

Margaret Graham
*margaret.graham@lw.com*
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

*/s/ Rose S. Lee*
**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

Rose S. Lee (*pro hac vice*)*
*roselee@mofo.com*
707 Wilshire Boulevard, Suite 6000
Los Angeles, CA 90017
Telephone: 213.892.5200

*/s/ Edward A. Bayley*
**KEKER, VAN NEST & PETERS LLP**
Robert A. Van Nest (*pro hac vice*)
*rvannest@keker.com*
Edward A. Bayley (*pro hac vice*)*
*ebayley@keker.com*
Paven Malhotra
*pmalhotra@keker.com*
Michelle S. Ybarra (*pro hac vice*)
*mybarra@keker.com*
633 Battery St.
San Francisco, CA 94111
Telephone: 415.391.5400

*Attorneys for OpenAI Defendants*

* All parties whose electronic signatures are included herein have consented to the filing of this document, as contemplated by Rule 8.5(b) of the Court's ECF Rules and Instructions.

**CERTIFICATE OF COMPLIANCE**

In accordance with Local Civil Rule 7.1(c), I certify that the foregoing Memorandum of Law is 5,502 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: December 10, 2025

*/s/ Margaret Graham*
Margaret Graham