

VIA ECF

January 30, 2026

Hon. Ona T. Wang
Southern District of New York
500 Pearl Street
New York, New York

In re OpenAI, Inc., Copyright Infringement Litigation, 25-md-3143 (SHS) (OTW)
This Document Relates To: 23-cv-8292; 23-cv-1021; 23-cv-11195; 24-cv-03285; 25-cv-04315

Dear Magistrate Judge Wang:

Since the first days of this litigation, Plaintiffs have been trying to understand which of their copyrighted works OpenAI copied, trained on, and output.¹ It is no secret to anyone – least of all OpenAI – that these are core questions in this MDL, and indeed a focus of more than two years of hard-fought, time-intensive, and expensive fact and expert discovery.

For years, OpenAI has consistently represented to both Plaintiffs and this Court that these questions are not easily answered, and that it does not have ready access to this information. OpenAI has resisted discovery into these topics by claiming (among other excuses): that it does not have this data, *see* Ex. A (OpenAI’s Response to Daily New Plaintiffs’ Interrogatory No. 1 refusing to identify instances of Plaintiffs’ content in the Training Datasets on the ground that doing so would require, *inter alia*, “OpenAI to conduct an analysis that Plaintiffs can conduct by expending the same effort and expense once provided with the relevant data.”); that it has no proportional way of searching for this data, *see* Ex. B (OpenAI’s responses to Class Plaintiffs’ RFAs stating that searching for specific copied data would require OpenAI to undertake an “unduly burdensome review of extremely burdensome data.”); that even if it did, mining such data poses serious user-privacy concerns that OpenAI takes very seriously, *see* Dkt. [841](#) at 1 (“The Order invades the privacy rights of millions of non-party ChatGPT users”); and that at bottom, it is just too burdensome – even in a MDL where OpenAI is alleged to have engaged in what may be the largest ever theft of copyrighted works.

Of particular note to this motion, many of OpenAI’s excuses were supported by declarations provided by Vinnie Monaco, a member of OpenAI’s technical staff. Dkt. [64-4](#), [435-1](#), [683](#), [1128](#).

But this past week, less than 24 hours before Mr. Monaco’s long-scheduled deposition, OpenAI produced a spreadsheet from its “NYT Recall Pipeline” analysis project (a part of its Project Giraffe, a project created for the purpose of identifying and attempting to block outputs

¹ Plaintiffs include The New York Times, *Daily News* Plaintiffs, *Ziff Davis* Plaintiffs, and Class Plaintiffs. Plaintiffs seek leave to exceed this Court’s 1,500 page limit in view of the late breaking discovery and where the parties were engaged in numerous depositions.

containing regurgitated training data). The spreadsheet shows that OpenAI’s representations were – at best – wrong. Indeed, the spreadsheet does exactly what Plaintiffs have been asking OpenAI to do for years: Identify instances where ChatGPT regurgitated a Plaintiff’s content in its output logs.² Ex. C. In other words, for the first time this week, OpenAI produced a spreadsheet identifying New York Times articles from its training data that appeared in its output logs – the product of a search that it previously claimed it could not do when Plaintiffs asked, but has in fact *already* done. By his deposition testimony, Mr. Monaco further confirmed that the process for blocking copyrighted material from appearing in outputs involves identifying that same material in the underlying training data.

In sum, it turns out OpenAI has long been searching its training data for copyrighted works, searching its users’ output logs to determine whether its models regurgitate copyrighted works, and then covering up its infringement by adding a filter to block the regurgitation in real-time. That information should have been produced years ago, and Plaintiffs may bring further motion practice addressing that delay and possible deletion. But for now, there is no question that what already exists should be immediately produced, and in no instance later than February 13.³

I. OpenAI Concealed Its Ability to Search Its Output Logs and Training Data

OpenAI engineer Mr. Monaco submitted four sworn declarations to this Court in support of OpenAI’s argument that creating a system to search existing consumer ChatGPT output logs for “n-grams” would be extremely burdensome – specifically, that it would require a “custom system” that “would take at least [REDACTED] of dedicated full-time engineer time” to design. Dkt. 64-4, ¶ 5. But Mr. Monaco’s deposition revealed that OpenAI had *already* built systems to search its training data for copyrighted content (including books and news content), search its users’ output log data, and identify instances in which its models output (or “regurgitated”) copyrighted content.

OpenAI withheld key evidence about these efforts until the day before Mr. Monaco’s deposition, when it produced a spreadsheet from its “NYT Recall Pipeline” analysis identifying instances where ChatGPT regurgitated New York Times content in de-identified ChatGPT conversations. Ex. C. Mr. Monaco confirmed that, in connection with what OpenAI called Project Giraffe, OpenAI had performed searches of deidentified ChatGPT conversation data for n-grams matching Plaintiffs’ content using a Python program that would perform this analysis. Indeed, according to Mr. Monaco’s contemporaneous documents, Mr. Monaco was an active participant in the project to add New York Times content to OpenAI’s anti-regurgitation filter to

² While OpenAI has only so far disclosed this spreadsheet for The Times’s content, it remains unclear whether spreadsheets exist for content owned by the other Plaintiffs. Even if they do not, the system OpenAI developed to create this spreadsheet for The Times’s content could likely have been used to search for other Plaintiffs’ content had OpenAI disclosed it had this capability.

³ As explained below, Plaintiffs also request six more deposition hours (not counting against any existing limits) with Mr. Monaco, who was woefully underprepared for his 30(b)(6) examination.

[REDACTED].” Ex. D; Ex. E, Rough 30(b)(1) Monaco Dep. Tr. at 11:36:27-11:42:17.

This process of searching output logs is precisely what OpenAI told both Plaintiffs and this Court it could not do. *Times* Dkt. 423 at 2 (“OpenAI does not have the preexisting technological functionality to fulfill Plaintiffs’ request to somehow preserve only user conversations that reference Plaintiffs’ content (to the extent all of Plaintiffs’ content could even be systematically identified)”). This is likewise precisely what led the parties, this Court, and Judge Stein to collectively invest hundreds of hours in the incredibly lengthy, burdensome, and expensive process of arriving at a sampling procedure for searching output logs so Plaintiffs could have some way of testing this element of their case. It turns out, that may not have been necessary. It certainly does not appear to be the only, or indeed the best, evidence in OpenAI’s possession on this critical piece of the case. Rather, OpenAI could search at least certain of its output logs for this information, and already had when it resisted providing this discovery in this case.

Turning to training, Plaintiffs moved incredibly early in this case for OpenAI to identify what works it trained on. *Authors Guild* Dkt. 78; *Times* Dkt. 305. OpenAI’s response was that it could not do so, resulting in this Court’s order requiring the construction of the “sandbox,” and the resulting hundreds of hours, and hundreds of thousands of dollars (at least), spent by Plaintiffs and their experts traveling to San Francisco where they would do their best to run searches for this very data in that restricted environment. But as Mr. Monaco testified at his deposition this week, the way OpenAI went about blocking Plaintiffs’ copyrighted works from appearing in ChatGPT outputs involved reaching back into the training data and first identifying the works there, so that they could then be blocked. Ex. E, Rough 30(b)(1) Monaco Dep. Tr. at 52:8-11 [REDACTED]

[REDACTED] What does this mean? Again, it means that OpenAI was not being accurate when it argued to the Court at the first discovery conference in September 2024 why it should not be compelled to identify what specific works were used in training. To the contrary, it had already done some of this work at that time.

As a result of OpenAI’s prior representations to News Plaintiffs and this Court, News Plaintiffs worked to secure the 20 million log sampling process. As this Court well knows, the process involved many months of negotiations, briefs, oral arguments, and confidential “settlement” negotiations, all of which took hundreds (if not thousands) of hours of attorney and expert time – to say nothing of the time this Court and Judge Stein spent on this issue, including addressing OpenAI’s vociferous complaints about user privacy. It now appears this may not have been necessary, because *OpenAI had the ability to search its output logs all along*. And its complaints about privacy and other issues aside, OpenAI had no problem searching its users’ output logs for this very information. Similarly, as a result of OpenAI’s representations, Plaintiffs have expended enormous resources in the sandbox, trying to identify training data that it turns out OpenAI may have already located.

Plaintiffs request that OpenAI be ordered to produce – by no later than February 13 – the Project Giraffe training data and deidentified output log datasets, source code, metrics, and

analyses related to OpenAI's efforts to detect and block its models from outputting copyrighted content contained in its pretraining datasets.

II. Mr. Monaco Was Unprepared to Testify About His Designated Topics

Mr. Monaco was designated by OpenAI on the following 30(b)(6) topics: (1) OpenAI's general techniques for limiting or preventing the regurgitation of text training data by the Relevant LLMs, including historical practices regarding the same (as to outputs), (2) OpenAI's policies governing the retention of user output data and technical mechanisms for implementing those policies, including efforts to preserve such data for purposes of the court's preservation order in this case, and (3) OpenAI's general practices regarding the use of data submitted by users of OpenAI's LLMs in the development of OpenAI's LLMs. As detailed below, Mr. Monaco was unprepared to testify on the topics for which he was designated, and Plaintiffs request an additional six hours not counting against time limits to conclude their questioning of Mr. Monaco, once he has been adequately prepared.

First, Mr. Monaco was unprepared on the topics for which he was designated. He testified that he only prepared for "a few" hours (which he then clarified was "between two and five" hours including interviews with "engineers."). *See* Ex. E (Rough 30(b)(1) Monaco Dep. Tr. Containing scores of examples of inability or unwillingness to answer questions that were plainly within the scope); Ex. F, Rough 30(b)(6) Monaco Dep. Tr. at 11:44:08 (Q [REDACTED])

id. at 11:23:06

Second, Mr. Monaco testified that he was a member of a Signal group in connection with his work at OpenAI. He testified that he received messages and he did not produce those messages in this litigation. Ex. E, Rough 30(b)(1) Monaco Dep. Tr. At 124:1-125:1. When asked why not, he speculated that they [REDACTED] *Id.* at 125:10-11. To Plaintiffs' knowledge, OpenAI has not collected, searched for, or produced Mr. Monaco's Signal chats. OpenAI should have produced these documents 14 days in advance under the deposition protocol and any future production after Mr. Monaco's deposition prejudices Plaintiffs. If the Signal messages at issue exist, they should be produced immediately. If they no longer exist, Plaintiffs are entitled to additional time to question Mr. Monaco on the circumstances of that deletion and the substance of the messages.

Third, despite having this data for years, OpenAI produced relevant documents and data concerning Giraffe two minutes before Mr. Monaco's deposition. *See, e.g.,* Ex. G (email from T. Gorman at 8:58 a.m. on the morning of the deposition noting the availability of Giraffe related data on the training data computer, among other documents relevant to the deposition produced that day); Ex. F at 92:2-7 (counsel noting that OpenAI's data management policy was produced

⁴ Plaintiffs also faced filibustering and time-wasting in the form of Mr. Monaco taking long pauses following nearly every question asked by News and Class counsel.

for the first time on the day of the deposition). Rather than beginning to take Mr. Monaco's 30(b)(1) testimony on the first day upon the conclusion of the 30(b)(6) portion, Plaintiffs were forced to defer to the next morning to allow them to examine the belated production materials. Even so, OpenAI's counsel halted the deposition in the middle of a line of questioning, asserting that Plaintiffs were "out of time" because the Parties had spent seven hours on the record. The deposition protocol in this case expressly provides for 11 hours of 30(b)(1) deposition time for fact witnesses questioned by both Class and News Plaintiffs. *See* Dkt. [355](#), at 3. That protocol instructs the parties to meet and confer in the event such a witness is also designated as 30(b)(6) witness. While there is no presumption of an 18-hour allowance under the protocol, OpenAI's arbitrary cutoff at the seven hour on day two mark is hardly a good faith limit, especially given that Plaintiffs were unable to use their full time on day one due to OpenAI's belated production and the issues stated above.

Plaintiffs therefore ask that Mr. Monaco's Signal messages be collected, searched and produced, and for six hours of additional time. Plaintiff further ask that this Court compel OpenAI to prepare Mr. Monaco by reviewing all documents asked about during his deposition, as well as datasets, source code, analyses, and metrics produced related to Giraffe.

January 30, 2026

Respectfully submitted,

/s/ Steven Liberman

Steven Lieberman

Rothwell, Figg, Ernst & Manbeck, P.C.

Counsel for The New York Times Company
and Daily News Plaintiffs

/s/ Davida Brook

Davida Brook

Susman Godfrey LLP

Counsel for The New York Times Company
Liaison Counsel for News Plaintiffs

/s/ Joshua Michelangelo Stein

Joshua Michelangelo Stein

Boies Schiller Flexner LLP

Counsel for Class Plaintiffs