February 13, 2026

In re OpenAI, Inc., Copyright Infringement Litigation, 25-md-3143 (SHS) (OTW)
This Document Relates To: All Actions

Dear Magistrate Judge Wang:

OpenAI confirmed three facts in its February 12 letter (Dkt. 1295).

First, since the outset of the litigation, OpenAI had in its possession reservoirs of 78 million and 10 million searchable output logs from a time period before OpenAI was using a filter to block output of Plaintiffs' content.

Second, since the outset of the litigation, OpenAI had tools to search these reservoirs of output logs, and in fact has conducted searches of these reservoirs for at least *The New York Times* and books content.

Third, since the outset of the litigation, OpenAI had tools to search its training datasets for Plaintiffs' content, and in fact has searched at least some of its training datasets for *The New York Times* and books content.

OpenAI should have – years ago – shared this information with Plaintiffs and the Court, and readily produced these materials. Plaintiffs request that the Court issue an order: (1) compelling OpenAI to immediately produce these materials and respond to the outstanding questions detailed below, (2) granting Plaintiffs' request for additional deposition time with Vinnie Monaco, and (3) awarding sanctions pursuant to Fed. R. Civ. P. 37(a)(5).[1]

OpenAI's letter is equally stunning for what it ***still*** does not adequately address, despite the ***years*** OpenAI has had to investigate these issues:

Question	OpenAI's Answer
Does OpenAI "have other such reservoirs, whether de-identified or not, of output logs that are searchable?" Ex. A at 9.	OpenAI does not address whether there are any other reservoirs of searchable output logs despite the two it (finally) disclosed to Plaintiffs. This is despite the fact that Mr. Monaco's testimony and documents produced in this case suggest other reservoirs of output logs exist. *See* Ex. Q at 3 (Feb. 2024 Slack thread stating [redacted]
Does OpenAI "have other searches that [it] already conducted for not just The New York Times' data, but The Daily News, CIR, class, so on, Ziff Davis, etc.?" *Id.*	OpenAI stated it is not aware of any other search for text that [redacted] or that [redacted] in "the collection" (presumably referring to the 78 million reservoir) or "[redacted]." Plaintiffs are concerned by how narrowly this is written. Regardless of how it was developed, which domain hosted the content,[2] or which OpenAI output logs it analyzed, has OpenAI identified text from Plaintiffs' works – or books

[1] Plaintiffs The New York Times, *Daily News* Plaintiffs, *Ziff Davis* Plaintiffs, The Center for Investigative Reporting, the Intercept, and Class Plaintiffs join in this motion.

[2] This is important because OpenAI has copied Plaintiffs' copyrighted works from non-U.S. websites known to plagiarize publisher content. Dkt. 72 ¶ 121 (OpenAI reproducing a NYT article obtained from dnyuz.com).

	and news content more broadly – in its output logs in any way aside from Monaco Exhibits 23 and 30?
Why didn't OpenAI "simply search [the output logs] that [it does] have access to for the copyrighted works," *id*., and why did OpenAI "lead [Plaintiffs] to believe they couldn't search [the output logs]" and "instead they had to do an inferior sampling process," *id*. at 14.	OpenAI does not address these questions other than to argue that the 78 million reservoir of output logs is "incomplete," cannot be used to conduct "statistical analyses," and production "would threaten the privacy" of its users. Dkt. 1295 at 2. It also suggests that ChatGPT output data compiled in May 2023 is deficient because it could contain "out-of-scope models," which makes no sense given that all models used to power ChatGPT in May 2023 are within the scope of this litigation.

I. Procedural Background

Plaintiffs have been trying to understand which of their copyrighted works OpenAI copied, trained on, and output since the first days of this litigation. After serving relevant requests in December 2023 and engaging in the usual meet and confer process thereafter, ***two years ago***, in February 2024, Class Plaintiffs moved this Court for OpenAI to admit whether their works appeared in OpenAI's training datasets. *Alter* Dkt. 78. At the September 12, 2024, Discovery Conference, OpenAI represented to this Court that it does "not currently have tools to search [its training datasets] efficiently." Ex. B at 7 (Sep. 12, 2024, Hearing Tr.). All Plaintiffs (and their various experts) then spent a year and a half setting up and visiting up the "sandbox" inspection environment in San Francisco, which continues to this day.

Fast forward another year, and after serving relevant requests in May 2024, ***more than a year ago***, in January 2025, News Plaintiffs moved this Court for OpenAI to preserve and produce data showing where their copyrighted works appeared in OpenAI's output logs. *Times* Dkt. 379. During the parties' meet-and-confer efforts, including those ordered by this Court, Plaintiffs asked whether the Project Giraffe tools may be used to identify their copyrighted works in OpenAI's output logs. Ex. C at 1. OpenAI represented that "the specific tools Plaintiffs identified did not function in a way that aligned with the request Plaintiffs were making." *Times* Dkt. 462 at 4. Based on OpenAI's representations, for the next year, the parties filed dozens of briefs, held three confidential settlement conferences with this Court, and engaged in extensive meet and confers, including likely thousands of emails and hundreds of calls that frequently included their technical consultants, to arrive at the 20 million sample of de-identified ChatGPT output logs.

OpenAI's prior representations to Plaintiffs and this Court about its abilities to search its training datasets and output logs proved to be false. Mr. Monaco confirmed that, under "Project Giraffe," [redacted][3], [redacted][4] and then [redacted]

[3] OpenAI ran searches over the training datasets for copyrighted content. Ex. E at 64, 82-83, 126-130; Ex. F; Ex. G; Ex. H [redacted]"); Ex. I [redacted] Ex. J [redacted]).

[4] Ex. K at 56-57; Ex. E at 51 [redacted]

[5] Ex. K at 98-99 ("[redacted]).

The spreadsheet produced the afternoon before the deposition of Vinnie Monaco, Dkt. 1220-1, Ex. D (excerpt of same, Monaco Exhibit 23) and source code produced two business days before the deposition[6] relate to a *New York Times*-specific evaluation under Project Giraffe referred to as the "NYT Recall Pipeline." OpenAI's February 12 letter confirms that the source code [redacted] " and that spreadsheet contains the [redacted]." Dkt. 1295 at 2. For example, as shown in the first example in the spreadsheet (Exhibit D), [redacted]

II. Plaintiffs' Requests for Production and Additional Deposition Time with Mr. Monaco

Mr. Monaco's deposition testimony identifies several gaps in OpenAI's February 12, 2026 letter to this Court and productions to Plaintiffs. Specifically, Mr. Monaco testified that OpenAI performed evaluations [redacted] Ex. K at 131-132. OpenAI has not produced the results of these evaluations or the reservoir(s) of output logs used for these evaluations. Instead, OpenAI's letter is narrowly focused on evaluations for: (1) a filter for content that [redacted] and (2) a filter [redacted] Dkt. 1295. Notably, when asked why he failed to mention [redacted] in any of his four declarations submitted to this Court, Mr. Monaco responded: [redacted] Ex. E at 52-53.

In view of OpenAI's deficient productions and insufficient explanations for them, Plaintiffs request OpenAI to produce:

(1) All of OpenAI's reservoirs of de-identified and searchable output logs, including but not limited to the 10 million and 78 million reservoirs, and any [redacted] datasets of de-identified output logs that can be rendered searchable;

(2) The results of OpenAI's searches for Plaintiffs' copyrighted content in its training datasets, to the extent not already encompassed by the "Giraffe" datasets;

(3) The results and statistical analyses of OpenAI's searches for copyrighted content from its training datasets in the reservoirs of output logs; and

(4) The tools OpenAI used for its searches and analyses of the output logs and training datasets, including the source code to the extent not already produced.

[6] The Protective Order restricts electronic dissemination of materials designated as source code, Dkt. 357 at 16, and the NYT Recall Pipeline source code was marked as Exhibit 22 during the deposition of Mr. Monaco. OpenAI previously clawed back the NYT Recall Pipeline code, later withdrew its assertion of privilege at the January 15, 2026 discovery conference, and delivered the code to Plaintiffs two business days before the deposition.

In addition, Plaintiffs request an additional six hours of 30(b)(6) deposition time with Mr. Monaco, not to count against Plaintiffs' hours caps, after the requested documents have been produced and Mr. Monaco has been adequately prepared to testify on those documents and the topics on which he was designated to testify, including: "OpenAI's general techniques for limiting or preventing the regurgitation of text training data by the Relevant LLMs, including historical practices regarding the same." Mr. Monaco could not answer foundational questions about OpenAI's efforts to block the output of verbatim copyrighted content. For example:

- Mr. Monaco could not answer or provide an estimate as to how many evaluations OpenAI performed on its filter to detect output of verbatim copyrighted content or the volume of ChatGPT conversation data used for these evaluations. Ex. K at 28-29.
- Mr. Monaco could not confirm whether any similar evaluations were done for API conversation data. *Id*. at 69.
- Beyond "[redacted]," Mr. Monaco could not identify what content OpenAI included in the filter. Ex. K at 39-42, 51-53.
- Mr. Monaco could not answer whether any of News Plaintiffs' content was included in the filter, other than [redacted] Ex. K at 44-45.
- Mr. Monaco could not confirm whether *New York Times* content was included in the filter even after being presented with a document stating [redacted] Ex. K at 48-50; Ex. M (Monaco Exhibit 2).

Mr. Monaco's lack of knowledge and evasiveness is inappropriate for a 30(b)(6) witness, and especially given OpenAI's reliance on his declarations to conceal their discovery deficiencies.[7]

III. OpenAI's Arguments Against Producing the 78M and 10M Reservoirs of Output Logs Fail

OpenAI's objections that (1) it "is not possible to use this collection of incomplete records to conduct the statistical analyses News Plaintiffs have articulated," and (2) it "would threaten the privacy of the tens of millions of ChatGPT users who generated the conversation data therein" should be rejected. Dkt. 1295 at 2. Plaintiffs requested a sample of 20 million logs for purposes of extrapolating over all ChatGPT output logs only because OpenAI represented that its output logs could not be searched without undue burden. Such sampling, however, does not excuse OpenAI from failing to search its reservoirs of de-identified and searchable output logs.

If OpenAI is aware of and has records of output logs with verbatim copyrighted content, those records should be produced. As confirmed by OpenAI's own documents and analysis, outputs in these reservoirs do appear to contain Plaintiffs' copyrighted content (Ex. D), and OpenAI found these records to be reliable enough to use to filter out copyrighted content in its ordinary course of business. In fact, these reservoirs of output log data may be some of the most probative ChatGPT data on questions around whether the models output Plaintiffs' copyrighted content because the reservoirs predate OpenAI's post-litigation efforts to block such outputs. Moreover, although OpenAI claims it is impossible to use these reservoirs "to conduct the statistical analyses News Plaintiffs have articulated," OpenAI fails to acknowledge the significant, ongoing issues about its deletions, substitutions, redactions, and myriad of other issues with respect

[7] Aside from these substantive problems, Mr. Monaco also took inexplicably long pauses throughout his deposition. *See, e.g.*, Monaco2 from 0:00-3:05 (12-18 second pauses when asked about the filter), Monaco2 from 1:23:52-1:27:02 (17 second, 21 second, and 27 second pauses when asked about data preservation efforts); Monaco 2 from 9:00-11:44 (12 second, 24 second, 20 second, and 21 second pauses when asked if Plaintiffs' content is in the filter).

to the 20 million log sample. Ex. A at 15 (recognizing that "we do have an ongoing issue about how and whether my May 13 order was complied with, given how explicit it was"). In fact, there was [redacted] user-initiated requests for deletions of ChatGPT output logs immediately following Mr. Altman's and OpenAI's public announcements on June 5, 2025, publicizing this Court's May 13 Preservation Order. Ex. K at 86-87; Ex. L at 3.

OpenAI's privacy arguments that "production of this collection would threaten the privacy" of its users are similarly misplaced because OpenAI already trains its models on these de-identified and PII-scrubbed reservoirs of ChatGPT data, and [redacted] Ex. N, Ex. O at 28-31 [redacted] This Court and Judge Stein have also repeatedly rejected OpenAI's privacy-related objections over the production of such data, in light of the robust de-identification process and protective order.

IV. This Court Should Grant the Requested Discovery and Award Sanctions

"Whether exercising its inherent power, or acting pursuant to Rule 37, a district court has wide discretion in sanctioning a party for discovery abuses." *Reilly v. Natwest Mkts. Grp. Inc.*, 181 F.3d 253, 267 (2d Cir. 1999); *see Design Strategy, Inc. v. Davis*, 469 F.3d 284, 294 (2d Cir. 2006). The imposition of Rule 37(a)(5) sanctions for failure to comply with discovery demands must be weighed in light of the full record. *See Cine Forty–Second St. Theatre Corp. v. Allied Artists Pictures*, 602 F.2d 1063, 1068 (2d Cir. 1979).

Plaintiffs are severely prejudiced by OpenAI's misrepresentations and belated disclosures about its abilities to search, and its prior searches over, its training datasets and output logs for Plaintiffs' copyrighted content. For the past year-and-a-half, News Plaintiffs' technical consultants alone have taken over 27 trips and spent over 400 hours for their onsite inspection of OpenAI's training datasets in San Francisco. They have also spent over 600 hours working on the output log sampling and review process. This is to say nothing of the attorney hours spent by all Plaintiffs' counsel meeting and conferring with OpenAI, crafting dozens of filings, and participating in three confidential settlement conferences with this Court related to sampling and discovery into OpenAI's output logs.

OpenAI's conduct was not the result of inadvertence. Among other things, its counsel was well aware that ChatGPT logs could be searched and ***had*** been searched – in fact, it claims that this was done "at the direction of legal counsel." Dkt. 1295 at 1. Yet it misled Plaintiffs and this Court by concealing this ability and misstating the burden associated with it. Accordingly, Plaintiffs request an opportunity to recover their fees and costs pursuant to Fed. R. Civ. P. 37(a)(5) related to the 20 million output log sample and onsite inspections in the training data sandbox. Plaintiffs may also seek substantive sanctions, including in the form of inferences and preclusions, after we complete our questioning of Mr. Monaco, finish discovery and analysis of all materials, and ascertain the full scope of prejudice.

Respectfully,

/s/ Steven Lieberman
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
News Plaintiffs' Liaison Counsel

/s/ Justin A. Nelson
Justin A. Nelson
Susman Godfrey LLP
Interim Lead Class Counsel

cc: All Counsel of Record (via ECF)