

VIA ECF

October 30, 2025
Hon. Ona T. Wang
Southern District of New York

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

Dear Magistrate Judge Wang:

News Plaintiffs submit this letter pursuant to this Court’s Order to further address Dkt. [656](#), News Plaintiffs’ motion to compel OpenAI to produce 20 million records of anonymized¹ consumer ChatGPT output log data from December 2022-November 2024 (the “historical” ChatGPT output logs). This is a simple motion to require production of evidence that is at the heart of this case: outputs of the models that OpenAI has trained and grounded without permission on Plaintiffs’ copyrighted works. OpenAI has emphasized the centrality of this evidence before this Court and in the court of public opinion, as it has repeatedly asserted that Plaintiffs “hacked” its products to produce examples of regurgitation, hallucination, and false attribution that threaten to destroy the market for original journalism. Now, after more than 500 days of negotiations, and in defiance of prior agreements, OpenAI seeks to hamstring Plaintiffs’ ability to put that very assertion to the test by refusing to produce even a small sample of the billions of model outputs that its conduct has put in issue in this case.

Immediate production of the output log sample is essential to stay on track for the February 26, 2026, discovery deadline.² OpenAI’s proposal to run searches on this small subset of its model outputs on Plaintiffs’ behalf is as inefficient as it is inadequate to allow Plaintiffs to fairly analyze how “real world” users interact with a core product at the center of this litigation.³ Plaintiffs cannot reasonably conduct expert analyses about how OpenAI’s models function in its core consumer-facing product, how retrieval augmented generation (“RAG”) functions to deliver news content, how consumers interact with that product, and the frequency of hallucinations without access to the model outputs themselves. Indeed, the interactions of Plaintiffs’ experts with the model outputs to perform this analysis are themselves work product to which OpenAI should not have access as a go-between.

Time is of the essence. To date, OpenAI has not produced any ChatGPT prompts or outputs to Plaintiffs. News Plaintiffs need adequate time to perform their expert analyses of these outputs, and to determine whether they will request additional sampling or productions (a right they expressly reserved when the parties limited the sample size to 20 million logs of the billions of existing ChatGPT logs). Dkt. [238](#). The fact that News Plaintiffs still do not have this data today is

¹ OpenAI has spent the last two and a half months processing and deidentifying this 20 million record sample, meaning that the communications in the logs are not traceable back to a specific individual or entity. In addition, this historical pool of 20 million logs does not include the temporary chats or user-initiated deleted chats that OpenAI contends are subject to heightened privacy protections because OpenAI already destroyed that data.

² It is also essential to assess potential issues of spoliation that are not yet before the court.

³ To the extent OpenAI contends some portion of these outputs are irrelevant, OpenAI and its experts can raise those arguments at summary judgment and/or trial.

particularly frustrating because it was OpenAI that proposed that it produce a 20 million sample—and News Plaintiffs agreed to dramatically narrow their requested sample of 120 million records in order to get quick access to this critical evidence so that they could begin their analytical work. Dkt. [394-4](#) (OpenAI’s June 25 letter offering “a **20 million sample size**” and rejecting Plaintiffs’ 120 million sample size because “anonymizing, decompressing, processing, producing, and analyzing such a large number of records under News Plaintiffs’ proposal would still impose a tremendous burden on OpenAI”); Dkt. [394-6](#) (OpenAI’s July 18 letter confirming it “offered to provide *millions* of conversations logs”); Dkt. [435](#) at 1 (arguing to “*first* proceed with OpenAI’s proffered 20 million log sample, and *then* have the parties meet-and-confer”); Ex. B (August 11 email). News Plaintiffs have been waiting months for OpenAI to anonymize this data, in order to ameliorate any concerns about user privacy. OpenAI should not be heard to claim that its own lengthy anonymization process is insufficient to mitigate the privacy concerns that it was designed to address.

OpenAI should be ordered to abide by its prior agreement and produce the 20 million log sample immediately.

A. The Parties’ Negotiations Over the Historical ChatGPT Output Log Data

The 20 million anonymized historical ChatGPT output logs represent just a tiny slice of the universe of output log data that News Plaintiffs have been requesting since May 2024. The figure below—which is not drawn to scale—provides a visual history as to how News Plaintiffs’ potential access to OpenAI’s output log data has been substantially narrowed through negotiation and this Court’s helpful involvement in the “discovery settlement” conferences.

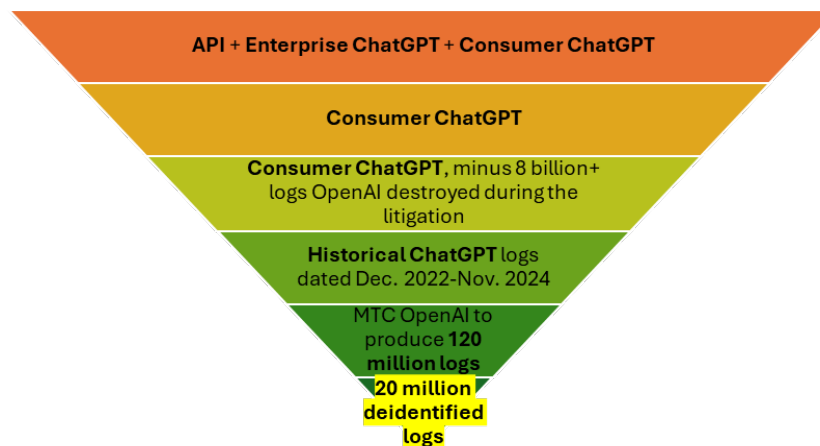


FIG. 1

On May 22, 2024, The Times served discovery requesting that OpenAI produce all output logs that contain or use News Plaintiffs’ copyrighted works. *See* Dkt. [311-2](#). These requests covered all of OpenAI’s products at issue in the litigation, including consumer ChatGPT, enterprise ChatGPT, and the Application Programming Interface (“API”). *See* FIG. 1, top row. While pursuing output log discovery, News Plaintiffs learned that OpenAI had destroyed all of the API logs and therefore cannot produce them. OpenAI also represented to News Plaintiffs (and this Court) that it does not have possession, custody, or control of the enterprise ChatGPT logs, and

thus could not produce those documents either. Ex. A (May 27, 2025, Hearing Tr.) at 22:2-9. Accordingly, only consumer ChatGPT logs were at that point potentially available for discovery purposes. *See* FIG. 1, second row. News Plaintiffs then learned that OpenAI had destroyed over 8 billion (and counting) consumer ChatGPT logs over the course of the litigation, which is the subject of other motion practice. Dkt. [212](#) at 3; *see* FIG. 1, third row.

This motion relates only to ChatGPT logs that users did not mark for deletion, which OpenAI retained in its ordinary course of business. For these “historical,” retained ChatGPT logs, the parties engaged in extensive meet-and-confers, including with their technical experts who offered differing views on an appropriate sampling methodology. Dkts. [394](#) and [435](#). On May 20, News Plaintiffs proposed a methodology to sample historical consumer ChatGPT data over a two-year period from Dec. 2022-Nov. 2024. *See* FIG. 1, fourth row; Dkt. [394-3](#).

The parties eventually reached an impasse as to the appropriate size of the sample to draw from the historical ChatGPT output log data. News Plaintiffs requested production of 120 million anonymized historical consumer ChatGPT output logs. *See* FIG. 1, fifth row. Because OpenAI would agree to produce only up to 20 million output logs, *see* Dkt. [394-4](#) (June 25 letter) and Dkt. [394-6](#) (July 18 letter), News Plaintiffs filed a motion to compel production of the full requested set of 120 million output logs (Dkt. [394](#)). *See* FIG. 1, sixth row. In opposing that motion, OpenAI announced, for the very first time, that it would take “roughly 12 weeks” for OpenAI to pull, decompress, and anonymize the sample of 20 million output logs. Dkt. [435](#) at 1. In view of OpenAI’s 12-week estimate, News Plaintiffs agreed to an initial set of approximately 20 million logs, while reserving their right to seek their “request for the full 120M samples.” Ex. B (August 11, 2025 email). OpenAI advised that it has been processing and anonymizing the requested 20 million logs for the past two-and-a-half months.

B. OpenAI’s Refusal to Produce the 20 Million Sample is Without Merit

On October 14, 2025, OpenAI announced that it was refusing to “produce” the 20 million historical ChatGPT output log sample because of “the privacy interests of OpenAI’s users.” Dkt. [656-1](#). For the very first time, OpenAI said that it will have to work with News Plaintiffs to “narrow[] the sample based on steps designed to identify the logs that are actually relevant to News Plaintiffs’ claims—including, for example, running searches to identify outputs with text regurgitated from news Plaintiffs’ articles.” *Id.* News Plaintiffs filed a motion to compel OpenAI to produce the agreed-upon 20 million sample of historical ChatGPT output log data. Dkt. [656](#). Instead of engaging in the Rule 26 analysis on burden and proportionality, OpenAI’s opposition (Dkt. [679](#)) reads like a press release to its customers. OpenAI’s purported concerns about user privacy are without merit for at least three reasons.

First, by the time the output log data samples will be ready for production, OpenAI will have spent three months pulling, decompressing, and anonymizing the records. According to OpenAI’s declarant Mr. Monaco, this process will “scrub categories of personally identifiable information and other information (*e.g.*, passwords or other sensitive information) from content like user data.” Dkt. [683](#), ¶ 3. Simply put, the output log data will not be tied to a specific individual or entity.

Second, there is a protective order in this case, and OpenAI may choose to designate the output log data with an appropriate level of confidentiality protections under the Protective Order. See, e.g., *Viacom Int'l Inc. v. Youtube Inc.*, 253 F.R.D. 256, 262 (S.D.N.Y. 2008); *Columbia Pictures Indus. v. Bunnell*, No. 06-cv-1093, 2007 WL 2080419, at *8 (C.D. Cal. May 29, 2007); *Simmons v. Danhauer & Assocs., LLC*, No. 08-cv-3819, 2009 WL 10677391, at *2 (D.S.C. Sept. 15, 2009). OpenAI's argument presupposes, without justification, that News Plaintiffs and their experts will violate that protective order.

Third, OpenAI's agreements with its customers provide that its use and disclosure of output log data is subject to its legal obligations.⁴ Moreover, OpenAI's agreement with its customers permits OpenAI to use output log data "to improve [its] Services, for example to train the models that power ChatGPT,"⁵ thus making clear to users that their prompts and the returned answers could be used for other purposes. In addition, the population of historical ChatGPT output log data does not have any heightened privacy considerations that OpenAI raised in connection with its oppositions to the Preservation Order because OpenAI has already destroyed the logs corresponding to temporary chats and user-initiated deletions.

In the most relevant case on this issue, *Concord Music Group, Inc. et al. v. Anthropic PBC*, the court ordered Anthropic to produce a 5 million prompt-output sample in a private discovery database walled off from Anthropic's counsel. Ex. C at 4 (*Concord Music Group, Inc. et al. v. Anthropic PBC*, Case No. 5:24-cv-03811-EKL, Dkt. 407 (N.D. Cal. Aug. 8, 2025)).⁶

C. OpenAI's Proposal to Narrow the Sample Unfairly Disadvantages News Plaintiffs

Instead of producing the sample of output logs so News Plaintiffs can review them, OpenAI wants News Plaintiffs to "work with" OpenAI's attorneys to somehow identify, without revealing News Plaintiffs' work product, what subset of the sample records—none of which News Plaintiffs have seen—are relevant to this case. This is not how discovery should work, and what OpenAI proposed would undermine the entire point of the agreed-upon sampling process.

First, the parties agree that the output log data is relevant—it's the best source of direct evidence to support News Plaintiffs' output-based infringement claims, and direct evidence of how "real world" users interact with a product that was built and operates by copying News Plaintiffs' content. In OpenAI's motion filed against The Times, OpenAI characterized such outputs as "not merely 'relevant' to The Times's copyright infringement claims," but that "they underlie essential elements of those claims and OpenAI's fair use defense." Dkt. 635 at 2. At the very outset of this case, OpenAI placed output log data at issue with respect to its fair use defense. *Times* Dkt. 52 at 1 ("In the real world, people do not use ChatGPT or any other OpenAI product" as a "substitute for a subscription to the New York Times."). OpenAI also continues to misstate the prevalence rate of infringement in the output log data. The 0.006% prevalence rate of infringement came from

⁴ See *Privacy Policy*, OpenAI, <https://openai.com/policies/row-privacy-policy/> (last visited Oct. 30, 2025) ("We may share your Personal Data . . . if required to do so to comply with a legal obligation").

⁵ See *Privacy Policy*, OpenAI at <https://openai.com/policies/row-privacy-policy/> (last visited Oct. 30, 2025).

⁶ In the case cited by OpenAI, the Court reduced a 2,500-sample set of conversations between customers and Noom weight loss coaches down to "the complaining subset" regarding subscription cancellations, and further noted that Noom "can elect to forego redaction and rely on the Protective Order." *Nichols v. Noom Inc.*, No. 20-cv-3677(LGS)(KHP), 2021 WL 1997542, at *4 (S.D.N.Y. May 18, 2021).

the *Concord v. Anthropic* litigation where the music publisher plaintiffs have accused Anthropic of outputting infringing copies of lyrics through its Claude chatbot. We currently do not know the prevalence rate of infringement, but OpenAI’s own public research on how people use ChatGPT suggests that 14-24% of all ChatGPT users search for information about people, current events, products, and recipes – namely, the very type of content the News Plaintiffs publish.⁷

Second, production of the 20 million record sample is proportional to the needs of the case. OpenAI has not identified any technical burden in producing the data, and it would be far less burdensome for OpenAI to produce the ChatGPT records than to have News Plaintiffs engage in protracted technical discussions and experimentation (at the risk of revealing their work product) over different mechanisms for analyzing the data. This is especially so for analyzing substantial, and hence infringing, similarity, which “presents one of the most difficult questions in copyright law, and one that is the least susceptible of helpful generalizations.” (Dkt. [701](#) at 4) (quoting *Structured Asset Sales, LLC v. Sheeran*, 120 F.4th 1066, 1078 (2d Cir. 2024)). No single search, or handful of searches, are likely going to capture the diversity of content in the millions of copyrighted works at issue or identify all infringing or substitutive outputs. Instead, News Plaintiffs’ work with the produced output logs will likely be an iterative process, with many searches built on the results of prior searches. Goldstein Decl., Dkt. [656-3](#), ¶ 4. News Plaintiffs will also wish to run searches over the metadata fields to analyze, for example, how their content was accessed and used as part of RAG.

It is improper for OpenAI to look over News Plaintiffs’ shoulders as they analyze the output log data, or to hamstring News Plaintiffs’ efforts by requiring burdensome onsite inspections that add weeks of delay. Goldstein Decl., Dkt. [656-3](#), ¶ 6. Unlike the training data that is magnitudes more voluminous, the 20 million anonymized consumer ChatGPT logs can likely fit on a single hard drive⁸ or can be readily produced in a secure cloud environment that is remotely accessible. Either option is acceptable to News Plaintiffs.

D. Intersection of the Historical ChatGPT Output Log Data Sampling with the Output Log Spoliation Issues

To be clear, deciding this motion does not in any way depend on issues related to spoliation. But prompt production of these 20 million logs will be relevant to the spoliation analysis, because it will more fully inform the prejudice⁹ that News Plaintiffs have incurred as a result of OpenAI’s deletion of billions of consumer ChatGPT output logs during the pendency of the litigation, and whether or how that prejudice may be cured.

⁷ Chatterji et al. “How People Use ChatGPT,” Sep. 15, 2025, pp. 2, 14 (available at <https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf>).

⁸ At Defendants’ request, Plaintiffs have produced data of a similar volume via hard drive, including deposit copies of the works at issue and voluminous metrics and website traffic data.

⁹ The spoliation sampling exercise has already demonstrated that News Plaintiffs have been prejudiced by OpenAI’s deletion of billions of historical ChatGPT output log data corresponding to temporary chats and user-initiated deleted chats during the pendency of the litigation. Dkt. [601](#). For example, users were far more likely to engage in ChatGPT prompts and outputs classified as news-related (and potentially paywall evading) as part of a “temporary chat” (Goldstein Decl., Dkt. [603](#), ¶8).

News Plaintiffs respectfully request this Court to order OpenAI to immediately produce the agreed-upon 20 million sample of anonymized historical consumer ChatGPT output log data either by: (a) putting that data on a hard drive and producing the hard drive; or (b) producing them to News Plaintiffs in a secure Cloud environment that is remotely accessible.

Respectfully,

/s/ Jennifer Maisel

Jennifer Maisel

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

*Counsel for The New York Times and Daily
News Plaintiffs*

cc: All Counsel of Record (via ECF)