ROTHWELL FIGG

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York

Re: *The New York Times Company v. Microsoft Corporation, et al.*, 23-cv-11195; *Daily News, L P, et al. v. Microsoft Corp., et al.*, 1:24-cv-3285

Dear Magistrate Judge Wang:

Pursuant to the Court's October 31 Order (Dkt. 304), The Times and Daily News Plaintiffs (collectively, the "News Plaintiffs") submit this joint letter with OpenAI to apprise the Court on the status of the "training data issues." News Plaintiffs have leave from OpenAI to file this letter.

***News Plaintiffs' Position**:* The identification of which of News Plaintiffs' copyrighted works Defendants took and used to train the GPT models (the ingestion of News Plaintiffs' content) is foundational to these cases and informs the scope of News Plaintiffs' claims and related discovery. But News Plaintiffs and OpenAI have a fundamental disagreement about who is responsible for identifying this information.[1] Because this information is solely within Defendants' possession, News Plaintiffs have served numerous discovery requests since February seeking information about the composition of OpenAI's training datasets and OpenAI's use of the News Plaintiffs' content to train the GPT models. *See, e.g.,* Ex. A (Rog No. 1); Ex. B (RFP Nos. 1, 8, 14, 20, 21, 22, 23); Ex. C (RFP Nos. 1, 3, 8, 14, 20, 21, 22, 23, 36, 93); Ex. D (Rog No. 10).

OpenAI responded to each of these requests with the following: "*OpenAI will make available for inspection, pursuant to an inspection protocol to be negotiated between the parties, the pretraining data for models used for ChatGPT that it locates after a reasonable search*." *Id.*[2] That is, instead of undertaking a search to identify responsive information in its datasets about its use of News Plaintiffs' works, OpenAI has insisted that News Plaintiffs must undertake their own inspection of OpenAI's datasets in a "very, very tightly controlled environment" that this Court has aptly referred to as a "sandbox." Sept. 12, 2024 Hearing Tr. at 5:23-6:6.

After months of negotiations, OpenAI finally agreed to an inspection protocol on October 4 (Dkt. 254)—which it stated was an express condition of News Plaintiffs beginning their inspection. News Plaintiffs (and their experts) noticed their inspection two business days later, and are now on their third week of inspection in the sandbox. The process has not gone smoothly, and they are running into a variety of obstacles to, and obstructions of, their review. These severe and repeated technical issues have made it impossible to effectively and efficiently search across OpenAI's training datasets in order to ascertain the full scope of OpenAI's infringement. In the first week of the inspection alone, Plaintiffs experienced nearly a dozen disruptions to the inspection environment, which resulted in many hours when News Plaintiffs had no access to the training datasets and no ability to run continuous searches. Ex. E. OpenAI has refused to install some software packages provided in advance by News Plaintiffs, delayed

[1] News Plaintiffs have produced more than sufficient information for OpenAI to identify the works at issue, including the copyright registrations as well as the titles, authors, URLs, and the full text contents of the articles.
[2] In response to meet and confers on News Plaintiffs' requests for production, counsel for OpenAI maintained that "OpenAI has responded to requests by making the training data available for inspection because it is in those training datasets where the information sought by the request can be found, if it exists." .

installing others, failed to properly install other software packages, and shut down ongoing searches that the News Plaintiffs had initiated. While News Plaintiffs are continuing to work with OpenAI to address technical issues that have come up, including bottlenecks due to the capabilities of the inspection environment, the process has been time consuming, burdensome, and hugely expensive. As of this date, News Plaintiffs have been forced to spend 27 person days (lawyers and experts) in the OpenAI sandbox and are nowhere near done. While they have already found millions of News Plaintiffs' works in the training datasets, they do not know how many more works are yet to be uncovered – information that OpenAI, as the party that chose to copy these works, should be ordered to provide.

In order to move forward with discovery, News Plaintiffs request the following relief from the Court: An order requiring OpenAI to identify and admit which of News Plaintiffs' works were used to train each of the GPT-1, GPT-2, GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, and GPT-4o families of models.

OpenAI below distorts the News Plaintiffs' request by focusing mostly on Requests for Admission that were recently served (Exs. F-G), but ignores that News Plaintiffs have sought the same discovery through numerous vehicles since discovery opened. While News Plaintiffs continue to bear the burden and expense of examining the training datasets, their requests with respect to the inspection environment would be significantly reduced if OpenAI admitted that they trained their models on all, or the vast majority, of News Plaintiffs' copyrighted content.

OpenAI's Position: There is no issue for the Court to decide here. *First*, Plaintiffs effectively ask the Court to *preemptively* compel responses to discovery requests that are not due for weeks. But until OpenAI serves its responses, there is no dispute. *Second*, their complaints about their inspection have either been resolved already, or—as Plaintiffs admit—they are the subject of ongoing discussions, so again there is nothing for the court to decide.

In the past week and a half, Plaintiffs served two sets of Requests for Admission. Exs. F-G. These requests demand separate responses for each "Asserted Work"—nearly 20 million in total—effectively resulting in almost 500 million requests. The deadline for OpenAI to respond is three weeks away. Fed. R. Civ. Pro. 36(a)(3). Plaintiffs chide OpenAI for focusing on their Requests for Admission, but yesterday—*after* the status hearing—Plaintiffs announced that they would seek unprecedented relief: an order compelling OpenAI to provide what it called "full responses" to these requests.[3] And Plaintiffs identify no other discovery requests that could possibly support an "order requiring OpenAI to … admit" facts.

The Federal Rules of Civil Procedure simply do not permit what Plaintiffs seek, which may explain why they cite no supporting rule or case. Under Rule 36(a)(6), Plaintiffs "may move to determine the sufficiency of an answer or objection" only after a response has been served, and Rule 37 permits a motion to compel a discovery response only after a failure to provide one. Here, OpenAI has not yet had a chance to serve its responses. Plaintiffs attach a

[3] At the status conference this week, Plaintiffs attempted to raise unspecified "training data sandbox issues," but the Court did not allow them to present their "one basic request" because it had not been briefed. Hearing Tr. (Oct. 30, 2024) at 63:24-65:6. The next day, Plaintiffs disclosed that their "one basic request" was for this preemptive order forcing OpenAI to answer RFAs before responses are due. Plaintiffs ***never*** **met and conferred** about this request.

mélange of other requests, but do not move for any additional relief (nor could they, as the parties have not met and conferred, the requests are unlinked to the relief they seek, or both).[4]

Taking a step back, everyone agrees the parties are navigating uncharted waters with training-data discovery. There are no precedents for such discovery, where Plaintiffs seek access to several hundred terabytes of unstructured textual data. OpenAI cannot easily identify the specific content that Plaintiffs are interested in, so it did exactly what Rule 34 allows: it invited Plaintiffs to inspect the data as it is kept in the ordinary course. There is no "sandbox." Rather, because the data is far too voluminous to produce, OpenAI built the hardware and software that Plaintiffs need to inspect. Specifically, OpenAI organized hundreds of terabytes of training data in an object-storage file system for Plaintiffs' exclusive use; it built an enterprise-grade virtual machine with the computing power to access, search, and analyze the datasets; it installed hundreds of software tools and tens of gigabytes of Plaintiffs' data upon their request; and it managed the necessary firewalls and secure virtual private network to support the inspection.

These efforts have enabled numerous inspections by Plaintiffs' counsel and consultants, including those from *Authors Guild et al., Raw Story Media, Inc., et al., In re OpenAI ChatGPT Litigation, New York Times Co., et al.,* and *Daily News LP, et al.* These inspections proceeded smoothly apart from the initial visit by consultants for the "News Plaintiffs." Based on information from Plaintiffs' consultants and OpenAI's investigation, it appears Plaintiffs' consultants triggered the technical issues they now complain about by overwhelming the file system with malformed searches. OpenAI has worked diligently to address these issues by providing on-call support, installing requested software, and then stress-testing the system. More importantly, after Plaintiffs—with OpenAI's guidance—implemented basic resource-management techniques, the disruptions ceased, and they apparently identified "millions of hits."

OpenAI will continue to assist Plaintiffs in overcoming technical challenges, provided Plaintiffs engage in good faith. Unfortunately, this has not always been the case. Although OpenAI first made training data available in early June, Plaintiffs scheduled their first inspection **only three weeks ago**. (Plaintiffs attribute this delay to how long it took them to agree to a protocol, but the protocol they agreed to in October is effectively the same as the one OpenAI offered months earlier.) While OpenAI agreed to install reasonably necessary tools, Plaintiffs submitted hundreds of irrelevant requests, including software for manipulating geospatial and audio data and software that does not exist. Two weeks ago, OpenAI offered to take over certain of the Plaintiffs' more defined searches—to lighten the burden on Plaintiffs' consultants—but Plaintiffs have inexplicably failed to provide any of information needed to move forward.

OpenAI remains willing to help Plaintiffs search the training datasets—not only by providing this unprecedented computing environment—but also by providing other technical assistance (it has already expended significant resources doing so). But Plaintiffs need to engage with OpenAI and meet and confer before seeking a court order compelling an answer to nearly 500 million requests for admission.

[4] For example, NYT RFP No. 36 seeks "Documents concerning the differences between Defendants' Generative AI models…," Daily News RFP No. 14 asks for "A complete listing of every data sample in the Training Datasets…," and NYT RFP No. 3 seeks "documents or communications regarding the appropriateness, value, importance, necessity, and/or concerns over or legal implications of the use of [copyrighted material]."

November 1, 2024

Respectfully submitted,

/s/ Steven Lieberman
Steven Lieberman

cc: All Counsel of Record (via ECF)