November 20, 2024

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:

I write on behalf of Plaintiffs The New York Times Company ("The Times") and Daily News, LP et al. ("Daily News") (collectively, the "News Plaintiffs") with a status update regarding training data issues identified at the October 30 conference and in the parties' joint letter to the Court dated November 1 (Dkt. 305),[1] and a renewed request that OpenAI be ordered to identify and admit which of News Plaintiffs' works it used to train each of its GPT models.

First, the News Plaintiffs continue to bear significant burden and expense in searching for their copyrighted works in OpenAI's training datasets within a tightly controlled environment that this Court and the parties have previously referred to as "the sandbox." OpenAI has provided the News Plaintiffs with two dedicated virtual machines with improved computing resources for performing their searches, and News Plaintiffs have spent an additional 150 person-hours (and even more computing hours) since November 1 searching OpenAI's training data. On November 14, all of News Plaintiffs' programs and search result data stored on one of the dedicated virtual machines was erased by OpenAI engineers. Maisel Decl. ¶ 3; Ex. A at 5. While OpenAI was able to recover much of the data that it erased, the folder structure and file names of the News Plaintiffs' work product have been irretrievably lost. Unfortunately, without the folder structure and original final names, the recovered data is unreliable and cannot be used to determine where the News Plaintiffs' copied articles were used to build Defendants' models. Maisel Decl. ¶ 4. Therefore, News Plaintiffs have been forced to recreate their work from scratch using significant person-hours and computer processing time. The News Plaintiffs learned only yesterday that the recovered data is unusable and that an entire week's worth of its experts' and lawyers' work must be re-done, which is why this supplemental letter is being filed today.[2] Maisel Decl. ¶¶ 4-5; Ex. A at 1-2.

Second, since the last hearing, the News Plaintiffs have sent OpenAI information for OpenAI to perform two separate searches on the News Plaintiffs' behalf: (i) on November 4, News Plaintiffs shared with OpenAI search terms corresponding to URLs that host, or have hosted, News Plaintiffs' content; and (ii) on November 13, the News Plaintiffs shared with OpenAI instructions for performing an "n-gram" search[3] in order to identify where portions of the News Plaintiffs' works appear in the training datasets. To date, the News Plaintiffs have not received results from either those searches, or confirmation that OpenAI has started them. On November 19, 2024, counsel for OpenAI reported only that they had "several promising meetings with OpenAI

[1] All docket references herein are to the docket in case no. 23-cv-11195.

[2] The News Plaintiffs provided OpenAI with an opportunity to file a joint submission, but they refused to do so. Maisel Decl. ¶ 5; Ex. A at 1-2.

[3] An "n-gram" refers to a phrase made of n-words, and the News Plaintiffs have proposed using 6-grams (6-word phrases) from the News Plaintiffs' works to run the search.

engineers recently." Ex. A at 1. On November 20, the same day as this filing, OpenAI served its responses and objections to the Daily News Plaintiffs' First Set of Requests for Admission, stating that it will "neither admit nor deny" whether Plaintiffs' works appear in the training datasets or were used to train the models. Ex. B.

* * *

The above developments, including OpenAI's erasure of a week's worth of work (which the News Plaintiffs have no reason to believe was intentional), underscore that OpenAI is in the best position to search its own datasets for the News Plaintiffs' works using its own tools and equipment. The News Plaintiffs have also provided the information that OpenAI needs to run those searches—all that is needed is for OpenAI to commit to doing so in a timely manner. Without such a commitment, the News Plaintiffs must reiterate their request, as set forth in the parties' joint November 1 letter, Dkt. 305, that this Court order OpenAI to identify and admit which of the News Plaintiffs' works it used to train each of the GPT models.

November 20, 2024

Respectfully submitted,

*/s/ Ian B. Crosby*
Ian B. Crosby

*/s/ Steven Lieberman*
Steven Lieberman

cc: All Counsel of Record (via ECF)