MORRISON FOERSTER

707 WILSHIRE BOULEVARD
SUITE 6000
LOS ANGELES
CALIFORNIA 90017-3543

TELEPHONE: 213.892.5200
FACSIMILE: 213.892.5454

WWW.MOFO.COM

MORRISON & FOERSTER LLP

AUSTIN, BEIJING, BERLIN, BOSTON, BRUSSELS, DENVER, HONG KONG, LONDON, LOS ANGELES, MIAMI, NEW YORK, PALO ALTO, SAN DIEGO, SAN FRANCISCO, SHANGHAI, SINGAPORE, TOKYO, WASHINGTON, D.C.

June 5, 2024

Writer's Direct Contact
+1 (213) 892-5656
ABennett@mofo.com

Via ECF

Hon. Sidney H. Stein
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007

Re: The New York Times Company v. Microsoft Corporation, et al.,
Case No.: 23-cv-11195-SHS: Discovery Dispute Regarding RFPs

Dear Honorable Judge Stein:

Since Plaintiff filed its letter brief, the parties have reached agreement on two of the three issues raised. Ex. A at 1. Those categories are discovery into the model training process (Plaintiff's Issue No. 2) and discovery about specific products (Plaintiff's Issue No. 3). *See* Mot. at 2-3 (discussing Issue Nos. 2 and 3); Ex. A at 1 (reaching agreement on those issues). The sole remaining issue for the Court is whether The Times is entitled to certain discovery related to three models that are not used for ChatGPT. It isn't.

In the spirit of cooperation, OpenAI has agreed to make available for inspection the training data for the text models identified in the complaint that are or have been made accessible through ChatGPT (GPT 3.5, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo), and the training data for two text models cited in the complaint that were never used to power ChatGPT (GPT-2 and GPT-3). Ex. A at 3. But, as explained below, models not used for ChatGPT have limited relevance to the issues in this case, and producing anything more than the training data that OpenAI has already agreed to make available for inspection would be disproportionate to the needs of the case. Fed. R. Civ. P 26. This is particularly true given that the models date as far back as 2018, and many of the employees who worked on them are no longer at the company.

Document Discovery Into GPT-2 & GPT-3 Is Disproportionate To The Needs Of the Case.

GPT-2 and GPT-3 were developed for research purposes in 2018 and 2019, respectively, and have not been used for ChatGPT. Given The Times's emphasis in its complaint on the "large

scale commercial exploitation of Times content" (ECF No. 1 ¶ 97)[1], OpenAI has to date focused its discovery efforts on the text models at issue in the complaint that were used for ChatGPT: GPT 3.5, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo. With the exception of GPT-4 Turbo (which was developed by further training on GPT-4), each of these models was trained from scratch,[2] meaning each was developed independently from the others and trained on distinct datasets. As a result, each model requires individualized, non-cumulative work to prepare its training data for inspection. For only the text models used for ChatGPT on or before the date the complaint was filed, that preparation has resulted in the collection of enough data to fill hundreds of hard drives—and required hundreds of hours of OpenAI employees' time.

Given the significant burden associated with making the training data for any model available for inspection and the limited relevance to The Times's claims of models not used for ChatGPT, OpenAI continues to believe that discovery associated with GPT-2 and GPT-3 is not proportional to the needs of the case. Fed. R. Civ. P. 26. To resolve this dispute, however, OpenAI has agreed to expand the scope of discovery to include inspection of the training data used to train GPT-2 and GPT-3. But the additional document discovery Plaintiffs seek has crossed the line of proportionality. Against the at-best limited relevance of these earlier models, Plaintiff's requests seeking, for example, "alternatives to using copyrighted content to train the Text Generation AI Models without compensation" cannot be justified. ECF No. 128-2 (RFP No.6). If upon inspection of the training datasets for GPT-2 and GPT-3, The Times identifies relevant information that it believes can justify further discovery into these models, OpenAI is willing to meet and confer about the scope of such discovery. But for the reasons given above, The Times is not entitled to such information without such a showing.

Discovery Into GPT Is Disproportionate to the Needs of the Case.

---

[1] *See also* ECF No. 1. ¶¶ 49-50 ("commercial purposes"), ¶ 63 ("commercial offerings"), ¶ 64 ("commercial success"), ¶ 65 ("OpenAI's widespread infringement commercial exploitation of Times Works"), *id.* ("design, development and commercialization of OpenAI's GPT-based products"; "widespread reproduction, distribution, and commercial use of Times Works"; "monetized the reproduction, distribution and commercial use of Times Works"), ¶ 74 ("creation and commercialization of the GPT models"; "massive copyright infringement, commercial exploitation, and misappropriation of The Times's intellectual property"), ¶ 102 (Defendants' commercial applications"); ¶ 156 (emphasizing "commercial uses" and "commercial purposes").

[2] *See* public statements at https://openai.com/index/approach-to-data-and-ai/. OpenAI is also prepared to submit a declaration from Nick Ryder, OpenAI's head of pre-training, confirming that the relevant models are trained from scratch, if such a declaration would be helpful to the Court. The Times challenges this assertion by citing a post (from an anonymous user) to a third-party internet forum. However, the post does not support The Times's position. At most, the post suggests that some of the GPT-3.5 training data may overlap with the training data from other models—an issue not in dispute and not relevant, since OpenAI is making available for inspection all the training data in its possession for any relevant model.

The Times also seeks discovery into OpenAI's GPT model. GPT was the first model released by OpenAI; it was released in 2018. Discovery regarding the GPT model suffers from the same proportionality concerns as GPT-2 and GPT-3, but there is even less support for relevance to the issues in this case. Plaintiff asserts (without evidence) in its letter motion that GPT is among the models alleged to have been trained on Times content. But The Times readily acknowledges that, like GPT-2, OpenAI released GPT on an open-source basis (ECF No. 1 ¶ 58). Unlike GPT-2, however, Plaintiff's complaint does not provide any reason to think that GPT trained on Times content. *Cf. id*. ¶¶ 85-87; *see also* Alec Radford et al., *Improving Language Understanding by Generative Pre-Training* at 4 (2018) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (explaining that GPT was trained on "7,000 unique unpublished books"). Accordingly, even if models not used for ChatGPT are deemed relevant to the issues in this case solely because they were allegedly trained on Times content, GPT is not among them.

In conclusion, we respectfully request the Court deny The Times's motion.

Sincerely,

/s/ Allyson R. Bennett

Allyson R. Bennett

Sincerely,

/s/ Michelle Ybarra

Michelle Ybarra
Keker, Van Nest & Peters LLP

Sincerely,

/s/ Andy Gass

Andy Gass
Latham & Watkins LLP

* The parties use electronic signatures with consent in accordance with Rule 8.5(b) of the Court's ECF Rules.