SUSMAN GODFREY L.L.P

June 11, 2024

Hon. Sidney H. Stein
United States District Judge
Southern District of New York
500 Pearl Street, Courtroom 23A
New York, New York 10007

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

Dear Judge Stein:

Plaintiff The New York Times Company ("The Times") respectfully requests a conference to address disputes with OpenAI related to The Times's First Set of "RFPs." The Times seeks an order compelling OpenAI to search for and produce five types of documents.[1]

1. OpenAI Refuses to Search for Policies Regarding Its Use of IP in Its Models.

RFP 5 seeks documents "reflecting policies, procedures, or practices concerning use of intellectual property in AI Models," including a policy known as "Copyright Shield" under which OpenAI agreed to "step in and defend [its] customers, and pay the costs incurred, if [they] face legal claims around copyright infringement."[2] Despite the plain language of the request and further meet-and-confers, OpenAI claims it is "unsure what documents you are looking for other than Copyright Shield." Ex. E at 5. OpenAI also has stated that it is willing to produce documents related to "opt-out" and "filtering," but will not otherwise conduct a search for responsive documents unless The Times identifies them. *Id.*

OpenAI's demand that The Times identify names of specific policies is improper. "Discovery is not to be a game of 'guess the right word,'" *Hollon v. Safeco Ins. Co.*, 2009 WL 10672218, at *2 (W.D. Okla. Jan. 21, 2009), nor may "defense counsel . . . transform discovery . . . into a game of hide-and-seek," *Charles v. City of N.Y.*, 2014 WL 1284975, at *14 (S.D.N.Y. Mar. 31, 2014). This RFP is clear by its plain terms, and "experience teaches that Defendant has some material which serves that purpose even if called by another name." *Hollon*, 2009 WL 10672218 at *2 (addressing RFP seeking "claims procedure manuals"). OpenAI should investigate and produce documents responsive to this request, from both custodial and non-custodial sources.

2. OpenAI Refuses Discovery into Alternatives to Using Copyrighted Content to Train Its Models.

RFP 6 seeks documents "concerning alternatives to using copyrighted content to train [Defendants' models] without compensation including using licensing agreements or material in

[1] The parties met and conferred by videoconference on May 6, and exchanged letters and emails before and after. The Times also filed a separate letter motion about other disputes for this set of RFPs. Dkt. 128. The Times's RFPs are attached as Exhibit A, OpenAI's Responses are attached as Exhibit B, and the parties' written correspondence is attached as Exhibits C, D, and E.

[2] *New Models and Developer Products Announced at DevDay*, OPENAI (Nov. 6, 2023), https://openai.com/index/new-models-and-developer-products-announced-at-devday/.

the public domain." Ex. B at 12. The only documents OpenAI has agreed to produce are agreements with third parties relating to training data (e.g., licensing agreements). *Id.*; Ex. E at 6. It refuses to produce communications about such agreements. It also refuses to produce other custodial documents concerning ways OpenAI could have trained its models without using copyrighted content (such as by using material in the public domain). Ex. D at 4; Ex. E at 6.

Those refusals are improper. If OpenAI considered, but decided against, training its models on non-copyrighted content, that is relevant to, among other things, The Times's claim of willful infringement and The Times's allegations that Defendants acted with the requisite scienter for purposes of the DMCA claims. *See* 17 U.S.C § 1202(b); Compl. ¶¶ 125, 187-90. Documents responsive to RFP 6 are also relevant to The Times's allegation that Defendants prioritized high-value content (like Times content) to train their models, as opposed to material that would have presented no intellectual property issues, such as works for which copyright protection has expired or works in the public domain. Compl. ¶¶ 87, 144-47. The only justification OpenAI has offered is that The Times refused to produce what OpenAI claims are similar documents related to licensing. But the parties are differently situated. The Times's communications regarding its license agreements have no relevance to this case, whereas OpenAI's use of copyrighted content for its models is core to the case. The Times appropriately seeks communications among OpenAI employees about alternative approaches to training its models and developing its products.

3. OpenAI Refuses to Produce Documents About Its Formation of a For-Profit Entity.

RFP 11 seeks documents "concerning OpenAI's transition to a for-profit company." Ex. B at 16. Initially, OpenAI sought to evade this RFP through word games, arguing that "no OpenAI entity has undergone" any "transition." Ex. D at 5. But when pressed during the meet and confer, OpenAI admitted that while no individual entity has "transitioned" from non-profit to for-profit status, OpenAI did ***create*** at least one for-profit entity. Ex. E at 3, 6. OpenAI nonetheless continues to refuse to produce documents responsive to this request.

This Court should order OpenAI to produce documents concerning its decision to form one or more for-profit entities. OpenAI's commercial designs bear on Defendants' anticipated fair use arguments, which include consideration of the "purpose and character of the use" of copyrighted works. *See A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001) ("This 'purpose and character' element also requires the district court to determine whether the allegedly infringing use is commercial or noncommercial."). OpenAI has suggested that discovery into this topic is unnecessary because it will not dispute that it charges fees to certain users of ChatGPT, but that does not come close to providing The Times the discovery it needs on this aspect of fair use. OpenAI touts that its work is intended to benefit all of humanity and it will undoubtedly press that narrative. *See* Dkt. 52-1 at 1. That OpenAI chose to create a for-profit entity through which it conducts most of its operations and solicited massive investments from Microsoft and others directly undermines that story. The Times is entitled to discovery on that choice.

4. OpenAI Refuses to Produce Documents About Collaborations with Microsoft.

RFP 15 seeks documents "concerning any commercial or supercomputing collaborations between and among the Defendants relating to Generative AI Models." Ex. B at 19. These documents are relevant to, *inter alia*, The Times's claim that OpenAI and Microsoft are joint infringers and The Times's contributory infringement claim against Microsoft. Compl. ¶¶ 93, 161-62, 174-77. OpenAI's refusal to produce any documents on the sole ground that Microsoft has already agreed to produce documents in its possession (Ex. E at 6) is meritless: both defendants have an obligation to provide relevant documents in their possession, and OpenAI undoubtedly has unique documents in its internal communications and files.

5. OpenAI Has Improperly Limited Discovery on Retrieval Augmented Generation.

RFPs 12 and 13 seek documents concerning Defendants' use of a technique called Retrieval Augmented Generation ("RAG"). RAG enables Defendants' products to integrate content from the "live" web—including The Times's website—with their large language models ("LLMs"). By using RAG, Defendants' products can generate answers to queries about current events and other information that postdates the training of the LLMs. These RAG-generated outputs display extensive excerpts or paraphrases of Times content—far beyond what a standard search result would look like. *See* Compl. ¶¶ 81, 108-23, 163, 179. The Times seeks documents concerning Defendants' use of RAG in their products, including documents about guidance on how particular websites are preferred or avoided in RAG results, and documents concerning how RAG search results are similar to or differ from standard search results. Ex. A at 8. OpenAI does not dispute the relevance of these requests.

Nevertheless, the only documents OpenAI has confirmed it will produce are those that are "sufficient to show how OpenAI uses RAG" and "sufficient to show OpenAI's understanding of RAG." Ex. E at 5; Ex. B at 17-18. That re-write of The Times's requests is improper. Defendants' use of RAG is a core aspect of The Times's claims, and a "sufficient-to-show" approach would exclude highly relevant documents, including communications among employees about whether Defendants should be using copyrighted content for RAG, as well as discussions about whether Defendants can and should employ safeguards to prevent their products from relying on copyrighted content. Any self-imposed "sufficient-to-show" limitation is improper because that approach "would let [OpenAI] decide what the answer[s] [are] on the merits [of factual issues in the case] and then limit discovery accordingly." *In re Stubhub Refund Litig.*, 2022 WL 1640304, at *2 (N.D. Cal. May 24, 2022).

OpenAI has evaded The Times's efforts to clarify whether it will produce any additional documents about this issue. When The Times asked if OpenAI would fully respond to these requests, OpenAI said it is "reviewing custodial documents that hit on RAG-related search terms" but refused to withdraw its previous sufficient-to-show limitation. Ex. E at 1. If OpenAI confirms in its response that it has withdrawn that limitation, then this dispute will be moot.

Respectfully submitted,

/s/ Ian B. Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

/s/ Steven Lieberman
Steven Lieberman
Rothwell, Figg, Ernst & Manbeck

cc: All Counsel of Record (via ECF)