VIA ECF

June 10, 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: No. 1:23-cv-11195 (SHS) (OTW)

Dear Magistrate Judge Wang:

I write on behalf of The New York Times Company and the *Daily News* Plaintiffs (collectively, "News Plaintiffs") concerning a dispute over OpenAI's failure to produce relevant datasets used to train or otherwise affect the performance of its Generative AI models. Those datasets—which contain copies of News Plaintiffs' content—are critically important and should have been produced more than a year ago. The requested datasets are listed in Exhibit A attached hereto and [redacted] and/or are referenced in OpenAI's source code and documentation relating to OpenAI's efforts: (a) to identify and rank high quality news content for training (a process sometimes referred to as "content scoring"), and (b) to suppress regurgitation of News Plaintiffs' copyrighted content. News Plaintiffs seek an order requiring OpenAI to search for and produce all datasets that are relevant to the News Cases.

I. Background

News Plaintiffs first served document requests on February 23, 2024–well over a year ago–seeking production of OpenAI's Training Datasets.[1] *See* Ex. B at 6 (RFP No. 1 seeking "[d]ocuments concerning the use of Times Content for training or fine tuning Your Text Generation AI Models"); *see also id*. at 7 (RFP No. 8). The Times's Second Set of RFPs to OpenAI seek production of Plaintiff-specific highly relevant datasets, including OpenAI's copy of the New York Times Annotated Corpus[2] and datasets that OpenAI compiled by scraping The Times's website. *See* Ex. C at 7 (RFP No. 74 which seeks "[a]ll documents relating to the NYT Datasets"); *see also id.* at 7-8 (RFP Nos. 77, 80-81). News Plaintiffs served additional discovery requests seeking datasets related to OpenAI's content scoring and quality ranking efforts for news content, including News Plaintiffs' copyrighted works. *See* Ex. D at 9, 10-11 (RFP Nos. 25 and 33). News Plaintiffs also served discovery requests seeking datasets related to regurgitation of News Plaintiffs' copyrighted works. *See* Ex. E at 11 (RFP Nos. 120 and 122).

Throughout News Plaintiffs' ongoing review of OpenAI's documents and source code, News Plaintiffs have identified several specific, responsive datasets—[redacted]—that have not been produced. *See* Exhibit A.

[1] News Plaintiffs' RFPs to OpenAI define "Training Datasets" as "the data and content used to pre-train, train, fine-tune, or otherwise affect the performance of the Generative AI Models[.]" *See* Ex. D at 4.

[2] The New York Times Annotated Corpus was a dataset containing over 1.8 million articles that The Times offered for public use under a strict non-commercial license. LDC_NYT_000033. The Times was able to confirm that OpenAI downloaded the corpus only after The Times subpoenaed and obtained documents from a third party, Linguistic Data Consortium, that was hosting the corpus. *See id.*

II. OpenAI's Production of Relevant Datasets is Incomplete

News Plaintiffs first identified specific missing datasets to OpenAI in a February 12, 2025 letter and in a March 25, 2025 follow up email. *See* Ex. F; Ex. G at 1-2. Since News Plaintiffs' March 25, 2025, correspondence, News Plaintiffs have sent three additional follow-up emails seeking to meet and confer with OpenAI on these issues, for a total of six requests, and have further identified additional responsive datasets that should have been produced. *See* Ex. H at 1-2; *see also* Ex. I. But it was not until June 4, 2025, when News Plaintiffs threatened to seek relief from this Court, that OpenAI agreed to meet with News Plaintiffs to discuss these issues. *See* Ex. H at 2-3.

During the parties' June 9, 2025 meet and confer, OpenAI said, for the first time in the four months since News Plaintiffs requested those materials, that some of the datasets listed in News Plaintiffs' February 12 and March 25 correspondences "might be privileged."[3] OpenAI did not explain how a dataset containing News Plaintiffs' content could possibly be privileged and even refused to confirm that it would produce these datasets if they are not privileged, saying that it needs to investigate the "relevance" of the identified datasets.

In April, OpenAI did produce a handful of datasets from News Plaintiffs' February list, but omitted several without explanation – including a dataset named [redacted] Moreover, some of the produced datasets were incomplete. *See* Ex. I. For example, OpenAI produced [redacted] versions of its News Plaintiff-specific anti-regurgitation datasets but has not produced versions of these datasets that have not been [redacted], as requested by News Plaintiffs in February. *See* Ex. F. Meanwhile, News Plaintiffs continued to do the work that OpenAI should have done in the first place, identifying an additional missing dataset in March 2025 [redacted] and—after further document review and inspection of OpenAI's source code—several more on June 4, 2025. *See* Ex. A at Nos. 14-17 [redacted] none of these datasets have been produced to News Plaintiffs.

III. OpenAI Should Produce the Missing Datasets

OpenAI has objected to producing the missing datasets on the ground that "[d]atasets that do not include [News Plaintiffs'] content are irrelevant and overbroad, and OpenAI is under no obligation to make them available." *See* Ex. H at 2 and 7. This objection makes no sense, given

[3] This issue is ripe for resolution despite OpenAI's belated privilege concerns expressed during the June 9 meet and confer. *See Barbera v. Grailed, LLC,* No. 24-CV-3535, 2024 WL 4836616, at *2 (S.D.N.Y. Nov. 20, 2024) ("[T]here is no loophole wherein a party may escape a motion to compel, or a ruling on such motion, simply by stonewalling its opposing counsel and declining to confer.").

[4] For avoidance of the doubt, it is News Plaintiffs' position that it should not be required to individually identify and request production of clearly responsive datasets. Instead, it is OpenAI's duty to identify and produce these responsive datasets in the first instance.

that OpenAI has refused to produce datasets that reference and include copies of News Plaintiffs' content.[5]

Moreover, the scope of relevant datasets to the News Cases is broader than datasets that contain verbatim copies of News Plaintiffs' works. For example, datasets relating to OpenAI's scoring or ranking of News Plaintiffs' content are relevant to News Plaintiffs' infringement claims and Defendants' fair use defense, regardless of whether these datasets include copies of News Plaintiffs' works. Indeed, News Plaintiffs have reason to believe that datasets stored at certain file locations identified in the March 25 email relate to [redacted]. These datasets are especially important in view of OpenAI's position that "any one sector—including news—is a tiny slice of overall training data, and any single data source—including The New York Times—is not significant for the model's intended learning." *See* Ex. J at 4.

Additionally, datasets related to anti-regurgitation and copyright infringement are relevant to News Plaintiffs' direct infringement claims and Defendants' fair use defense. For example, News Plaintiffs have identified a dataset titled [redacted] that News Plaintiffs believe contains [redacted]. *See also* Ex. A at Nos. 2 and 6 (additional regurgitation datasets not produced by OpenAI). These datasets—many of which were created after News Plaintiffs put OpenAI on notice of infringement—are particularly relevant in view of Defendants' position that regurgitation is an "uncommon and unintended phenomena." *See Times* Dkt. 52 at 10; *see also Daily News* Dkt 82 at 6. OpenAI's post-complaint anti-regurgitation datasets are particularly important in view of the parties' ongoing issues concerning OpenAI's deletion of output log data.

IV. OpenAI has not articulated an undue burden in producing all relevant datasets

The requested datasets are readily identifiable and OpenAI has not articulated an undue burden associated with producing the missing datasets identified by News Plaintiffs in accordance with the parties' negotiated training data and ESI protocols. With respect to the content scoring datasets, OpenAI has asserted only that the directories containing these datasets are "enormous" and thus inspection of these datasets is not practical. Ex. I at 2. However, "sheer volume alone is an insufficient reason to deny discovery of documents," and the parties have already been operating for nearly eight months under a custom training data inspection protocol designed to address production of large datasets. *In re Adelphia Commc'ns Corp.*, 338 B.R. 546, 553 (Bankr. S.D.N.Y. 2005); *Leibovitz v. City of New York*, 2017 WL 462515, at *2 (S.D.N.Y. Feb. 3, 2017).

Because the datasets sought are relevant and proportional to the claims in this case, News Plaintiffs ask that this Court order OpenAI to produce: (1) the datasets identified in Exhibit A, and (2) directory information for the following dataset locations: [redacted] so that News Plaintiffs may identify the complete group of missing datasets.

[5] News Plaintiffs' correspondences with OpenAI have additionally identified a dataset [redacted] that News Plaintiffs believe includes their content. OpenAI has neither confirmed nor denied whether this is the case.

Respectfully,

/s/ Ian Crosby
Ian B. Crosby
Susman Godfrey LLP

/s/ Steven Lieberman
Steven Lieberman
Rothwell Figg Ernst & Manbeck, P.C.

cc: All Counsel of Record (via ECF)