KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

August 4, 2025

VIA ECF
Hon. Ona T. Wang
U.S. District Judge for the Southern District of New York
Daniel Patrick Moynihan United States Courthouse
500 Pearl Street
New York, NY 10007

RE: *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143, This Document Relates To: 23-cv-11195

Dear Judge Wang:

OpenAI submits this letter in response to News Plaintiffs' extraordinary request that OpenAI produce the individual log files of 120 million ChatGPT consumer conversations. ECF 394. This request is inappropriate for two primary reasons.

First, Plaintiffs vastly underestimate the burden and time required to generate a de-identified sample of 120 million records. *Infra* § 1. Generating these samples is a multi-step process of retrieval, decompression, processing, de-identification, and storage. The de-identification process, *e.g.*, will require passing each sample through OpenAI's custom de-identification tool [redacted] *see* Monaco Decl. ¶12). That means using a de-identification pipeline that OpenAI already uses for other business-critical functions (which will take longer for a larger sample) or, in the alternative, building an entirely separate pipeline (which will take months to develop). Monaco Decl. ¶¶8–11. That means either interfering with critical business functions or building an entirely separate pipeline for just this single piece of this process. As such, increasing the size of the sample from 20 to 120 million will ***vastly amplify*** both the cost of this exercise (from roughly $[redacted] to roughly $[redacted]), as well as the time it will require (from roughly 12 weeks to up to 36 weeks). The lengthy period of time that such processing requires would threaten to derail the case schedule. The Court should not order OpenAI to undertake that drastic step absent a compelling showing of need, which Plaintiffs have not presented.

Second, OpenAI has in fact proffered a ***20 million*** conversation sample. This size is surely more than enough to conduct appropriate analyses relevant to Plaintiffs' claims. *Infra* § 2. If Plaintiffs cannot find evidence that ChatGPT has regurgitated their articles in 20 million conversation logs, that illustrates that ChatGPT does not do so with any meaningful frequency. Plaintiffs do not meaningfully contest that point. Instead, Plaintiffs insist that they should be entitled to conduct a full-scale analysis on every single month during the relevant 23-month time period—notwithstanding the burden—so that they can evaluate how the product has changed over time. But that kind of extraordinarily granular analysis is disproportionate to the issues in dispute. Further, Plaintiffs have not proffered a single reason why a 20 million sample is in fact

insufficient for their claims in this case. The only expert who has examined this precise question found that a 20 million sample data set would be more than sufficient to analyze the differences in outputs over time. *See* Berg-Kirkpatrick Decl. ¶15.

Simply put, Plaintiffs' demand for 120 million individual conversation logs would impose a massive burden on OpenAI, delay this case by months, increase the scope of user privacy concerns, and yield no discernible statistical benefit. The Court should deny it. At minimum, in light of the burdens at issue, the Court should proceed incrementally by ordering the parties to *first* proceed with OpenAI's proffered 20 million log sample, and *then* have the parties meet-and-confer if News Plaintiffs can demonstrate that their ability to prosecute their claims will be materially prejudiced absent another targeted sample.

1. Plaintiffs' Proposal Would Vastly Increase Burden and Prolong the Case Schedule

Plaintiffs seek 120 million records from OpenAI's offline storage system, which is composed of individual conversation logs.[1] The logs are not rows in a spreadsheet; they are large, unstructured data files—meaning that they do not follow a predefined format—consisting of over 5,000 words, even for very short conversations. Berg-Kirkpatrick Decl. ¶16. The logs must be decompressed before being searched and contain identifying information (*e.g.*, addresses) and other private information (*e.g.*, passwords) that must be scrubbed before making it available. Monaco Decl.¶6–8.

The process of generating a sample of these logs is highly complex and is in no way comparable to the process of running 800-word excerpts through an off-the-shelf tool. *Contra* ECF 394-1 at 3. Generating this sample will require OpenAI to (i) retrieve each requested log by finding it in the tens of billions of logs in OpenAI's offline data storage, *see id.* ¶4–5; (ii) decompress each log, *id.* ¶6; and (iii) parse each log, including by sorting through layout changes that occurred during the Study Period, *id.* ¶7. The logs must then be (iv) de-identified using the same rigorous process that OpenAI applies to other user data. *Id.* ¶8. That requires application of [redacted]. Monaco Decl. ¶8, 12. That means balancing this process against other business-critical operations that rely on the same pipeline (which constrains OpenAI's ability to process the logs quickly) or, in the alternative, building an entirely separate pipeline (which will take months to develop). *Id.* ¶9–11. The logs must then be (v) stored, which adds to the costs. *Id.* ¶13. Each of these steps requires time, computational resources, and OpenAI engineers to design, debug, operate, and monitor the relevant systems.

Many of these costs—including the costs of retrieving, parsing, and de-identifying the data—are variable and increase linearly with sample size, which means that Plaintiffs' proposal can be up to 6 times as expensive (and time consuming) as OpenAI's proposal. *Id.* ¶5, 7, 10. For that reason, while OpenAI estimates that it could generate a 20 million conversation sample in roughly 12 weeks (costing roughly $[redacted]), generating a 120 million conversation could take up to 36 weeks (costing up to $[redacted]). *Id.* ¶3.

[1] The data source for this merits sampling exercise is distinct from the data source for the proposed sampling protocol for preserved data. 2025.05.27 Aft. Tr. at 36:20–37:5.

Plaintiffs' proposal, in other words, would massively increase both the financial burden of this exercise and the time required to generate a sample, delaying expert analysis and threatening the litigation timeline. The Court should not order that drastic step absent a particularly compelling showing of need, which Plaintiffs have not presented here.

2. Twenty Million Conversations Are More Than Sufficient

Plaintiffs do not dispute that 20 million conversation logs are more than sufficient to determine prevalence rates of alleged infringement—that is, how often infringement may have occurred during the relevant time period of "December 2022 through November 2024" (the "Study Period"). ECF 394 at 1. Put simply, if Plaintiffs cannot find evidence of infringing outputs in 20 million conversation logs, that shows that ChatGPT does not produce such outputs in any meaningful frequency. *Cf. Authors Guild v. Google, Inc.*, 804 F.3d 202, 223 (2d Cir. 2015).

Nonetheless, perhaps recognizing that the kinds of outputs they are searching for are vanishingly rare, Plaintiffs insist on a 120 million sample—*i.e.*, a sample that is ***twenty four times larger*** than the samples that courts have ordered in other similar cases. *Concord Music Group, Inc. v. Anthropic PBC*, No. 5:24-cv-03811, ECF 377 (N.D. Cal. May 23, 2025). Plaintiffs' only justification for their request is a desire to slice-and-dice the relevant 23-month period into months and conduct full-scale analyses on each month's data so that they can see how, *e.g.*, the prevalence of regurgitation changed over time. But that kind of granular, zoomed-in analysis is unnecessary to evaluate the prevalence of regurgitated outputs during the relevant period. Plaintiffs do not explain why this over-time analysis is relevant to the merits. They do not cite any authorities in support of their approach. And they do not explain why the probative value of their over-time analysis outweighs the burden and time required to generate their vastly expanded sample.

More to the point, neither Plaintiffs' Motion nor their supporting expert declaration contains any statistical analysis suggesting that Plaintiffs need more than 20 million samples to conduct that kind of over-time study. That is because they do not: as Dr. Berg-Kirkpatrick explains in his attached declaration, 20 million conversation logs are more than sufficient to draw statistically valid conclusions about differences within narrow windows within the Study Period. *See* ECF 394 at 1-2; *see* Berg-Kirkpatrick Decl. ¶15. That sample size will, *e.g.*, enable Plaintiffs to perform their requested analysis: study the prevalence of regurgitated outputs in millions of sampled conversations from "the March 2023–July 2023 time period" and compare that result to the prevalence of such outputs in millions of samples conversations from "the July 23, 2023–September 2023 time period." ECF 394.

3. Plaintiffs' Request Is Premature

The Court should therefore deny Plaintiffs' request for 120 million logs. In the alternative, the Court should proceed with 20 million logs unless and until News Plaintiffs can demonstrate that their ability to prosecute their claims will be materially prejudiced absent another sample. In that event, the parties can confer about a targeted supplemental sampling to address Plaintiffs' specific, demonstrated need.

Respectfully submitted,

KEKER, VAN NEST & PETERS LLP	LATHAM & WATKINS LLP	MORRISON & FOERSTER LLP
/s/ Edward A. Bayley	*/s/ Elana Nightingale Dawson*	*/s/ Rose S. Lee*
Edward A. Bayley*	Elana Nightingale Dawson	Rose S. Lee*

cc: All Counsel of Record (via ECF)

* All parties whose electronic signatures are included herein have consented to the filing of this document.