KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

November 24, 2025 **VIA ECF**

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

Re: Contested Download and Output Discovery, *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143; This Document Relates To: *Baldacci*, No. 1:25-md-3143-SHS-OTW; *Authors Guild*, 1:23-cv-08292-SHS-OTW

Dear Judge Wang:

OpenAI writes in response to the Court's November 13 Order (ECF 751) directing the parties to confer on any further discovery arising from Judge Stein's recent Orders (ECF 701 and 707) and to address any disputes as to the burden or proportionality of Plaintiffs' requests.

Because OpenAI has already provided the vast majority of discovery that Plaintiffs could need to support their download and output claims, the parties should not need (and are not currently requesting) any change to the February 27, 2026 fact discovery cutoff. Indeed, Plaintiffs represented to the Court in May that they already "have the facts" to assert the download claim and could "get certainly to a resolution" on the claim "with literally a couple months more of discovery." 5/22/2025 Hearing Tr. 52:2-7. The same is true for Plaintiffs' output claim. Plaintiffs have taken and received discovery relevant to their output claim.

Unfortunately, Plaintiffs' positions in recent meet and confer are not consistent with their representations to the Court or with completing fact discovery on time. For instance, as the Court is aware, Plaintiffs' download claim focuses on the allegation that OpenAI has improperly downloaded books from so-called shadow libraries. OpenAI has already provided extensive discovery into how such books were acquired, including discovery on the one instance in which an employee used a torrenting protocol. Plaintiffs now want full discovery into *any* instance in which *anyone* used a torrenting protocol, even if the file had nothing to do with a book.

Plaintiffs' "boil the ocean" approach is not proportional to the needs of the case and risks an adjustment to the case schedule. Without guardrails, Plaintiffs' overbroad demands will mire the parties in many months of tangential and irrelevant discovery and delay a ruling on merits issues of public interest. OpenAI has thus offered Plaintiffs an alternative approach, as discussed below, that provides the additional discovery they need while preserving the schedule.

Download Claim Discovery

Plaintiffs previously represented in meet and confer that their priority for their download claim is to ensure that OpenAI has identified and made available copies of books OpenAI downloaded from specific shadow libraries, as well as related communications. OpenAI has already done so. While OpenAI is not aware of any additional non-duplicative documents on this point, it has agreed to perform additional reasonable searches to confirm this.[1]

Now, contrary to their representation that they already "have the facts," Plaintiffs demand (1) discovery into any torrenting of non-book data; (2) searches of billions of (likely irrelevant) files in Azure containers; (3) access settings for any files sourced from a shadow library, regardless of burden; (4) production of each and every "copy" of book file downloads; and (5) additional technical documents. For each of these disputes, OpenAI has already conducted or is proposing a reasonable search.

Torrenting Discovery

Plaintiffs have demanded that OpenAI undertake an extensive investigation to learn if *anyone* at the company *ever* used a torrenting protocol to download *any* data, at any point. Beyond that, Plaintiffs demand detailed technical information about each such download. Plaintiffs *already have* this discovery for books data. In response to requests served nearly a year ago, OpenAI has spent hundreds of hours of attorney and employee time searching for documents and data related to the use of the BitTorrent protocol to acquire data from Library Genesis. OpenAI has already produced all the evidence it is aware of on the single known episode, including source code, configuration files, Slack messages, invoices, cost reports, and so on. Plaintiffs long ago agreed that this discovery satisfied OpenAI's obligations on torrenting RFPs.

Now, and absent any nexus to Judge Stein's decision on OpenAI's motion to strike, Plaintiffs demand similar discovery into torrenting of *other*, ***non-book*** data. To what end? In conferral, Plaintiffs hypothesized that the discovery *might* show that a different employee used different torrenting settings for uncopyrighted data. This is plainly irrelevant. Plaintiffs are demanding an investigation that could take months to complete and is untethered to any party's claims or defenses. Class Plaintiffs are book authors, and their claims relate to copyrighted books. *See* ECF 183 (Consol. Compl.) ¶ 294 (defining proposed class to include "legal and beneficial owners of copyrighted works"). Fishing expeditions into other types of irrelevant data are disproportionate to the needs of the case.

[1] Shortly before the filing deadline, Plaintiffs informed OpenAI of their plan to raise non-responsive "hits" to specific search terms associated with "Library Genesis." OpenAI recently conferred with Plaintiffs on this subject, answered Plaintiffs' questions, and explained that, while it had no reason to doubt its prior determinations, it would look again at those documents in a good faith effort to resolve the dispute.

Azure Containers

As part of its search for books purposefully downloaded from shadow libraries, Plaintiffs demand that OpenAI search Azure containers associated with 22 employees. Those containers contain over ***15 terabytes of data*** and nearly ***4 billion files***. If OpenAI had reason to believe an employee stored raw book files downloaded from Plaintiffs' specified list of "shadow libraries" in a personal Azure container, OpenAI collected this data and produced it. OpenAI researchers testified that they do not have a systematic way to search the containers given their size and technical structure. Nevertheless, OpenAI is exploring solutions that would allow for a reasonable search to identify data downloaded from Plaintiffs' specified sites, if any.

Plaintiffs, however, demand that OpenAI apply search terms across some containers and let Plaintiffs inspect directory names for others, without regard to relevance. Neither approach is workable. Given the technical structure of the containers, and the enormous volume of data stored there, applying search terms to the underlying data is not feasible. OpenAI estimates that the required processing work could cost tens of millions of dollars. Similarly, production of directory names would involve production of *millions* of folder names without regard to relevance. Given that the vast majority of stored data in personal containers (if not all of it) is irrelevant to Plaintiffs' claims, Plaintiffs' proposal is burdensome and overbroad. Instead, Plaintiffs should let OpenAI continue its work, and the parties should continue to confer.

Access Settings For Technical Systems

Plaintiffs seek documents sufficient to show which employees had access to books downloaded from Library Genesis or other shadow libraries (if any). OpenAI has proposed a reasonable solution that Plaintiffs have accepted, which is to look for and produce the applicable data access/security policies from the relevant time period. That should resolve this dispute.

But Plaintiffs want a commitment now to undertake far more burdensome searches in the event OpenAI cannot locate policy documents. That demand is premature. OpenAI should undertake the search it is proposing and if it cannot locate the relevant documents, the parties can continue their conferral about a reasonable alternative source of the information Plaintiffs seek.

Discovery Into Every "Copy"

There is a reason Plaintiffs were able to report that they already "have the facts" to assert the download claim. As noted above, OpenAI has either already produced all of the "raw" books data downloaded from Plaintiffs' list of shadow libraries or agreed to perform a reasonable search to confirm it has done so. At the other end of the training pipeline, OpenAI has produced processed training datasets for the relevant models. Relatedly, OpenAI also has produced relevant source code that explains how each step in OpenAI's model training process works.

With their request for discovery into "copies" and "uses" of books downloaded from Plaintiffs' list of shadow libraries, Plaintiffs appear to be asking for discovery into intermediate steps of the LLM training pipeline. For example, text extraction, deduplication, filtering, and tokenization create new, distinct datasets that are part of the process for training a model. If OpenAI understands Plaintiffs' request, they seek these "intermediate" steps that are known to contain books (or parts of books) downloaded from the specified shadow libraries. While far from clear that demanding "copies" is the best way to obtain discovery about the training pipeline (as opposed, for example, to asking witnesses at deposition), OpenAI is prepared to perform a reasonable search to satisfy this request.

Technical Documents for Other Books Data

Two hours before the filing deadline, Plaintiffs clarified they seek even more technical documents. As to books datasets downloaded from non-shadow library sources, OpenAI's ESI searches were broad enough to capture relevant documents and OpenAI produced them. OpenAI is open to producing the datasets themselves, subject to relevance and proportionality considerations. But the parties should confer about what additional datasets Plaintiffs are requesting. As to source code, OpenAI's understanding is Plaintiffs have this already.

As to Plaintiffs' demand relating to a "book summarization" project, OpenAI agreed to conduct a reasonable ESI search. Plaintiffs accepted this but, hours before filing, clarified their demand extends to technical documents. OpenAI will produce non-privileged relevant technical documents where proportional to the needs of the case, but needs time to assess these issues.

Output Claim Discovery

OpenAI also has already provided discovery related to Plaintiffs' output claim and agreed to provide more, such as by (1) running targeted searches, (2) allowing Plaintiffs to inspect the relevant models, and (3) offering additional document discovery. All of that can be accomplished by the close of fact discovery. As with download discovery, however, Plaintiffs are overreaching and demanding more than is proportional to the needs of this case.

ChatGPT Output Logs

OpenAI has offered to conduct n-gram searches and targeted keyword searches sufficient to identify outputs related to Plaintiffs' asserted works and to confer with Plaintiffs on search terms. This proposal will give Plaintiffs the output discovery they need without unnecessarily invading the privacy of millions of ChatGPT users.

Without explaining why the offer is insufficient, Plaintiffs demand production of the 20 million private ChatGPT user conversations the Court recently ordered OpenAI to produce in the News cases. OpenAI respectfully objects to that production for the same reasons set forth in its prior briefing and its Rule 72 objection. *See, e.g.*, ECF 681, 683, 717, 742, 752, 840, and 841.

Disclosure of these private conversations to Class Plaintiffs–without notice, consent, or any showing of relevance to the infringement claims–would exacerbate the privacy concerns OpenAI has already identified.

Output Document Discovery

Plaintiffs also demand that OpenAI search for additional documents related to the use of LLMs for commercial writing, including broad sub-categories such as "user behavior" studies, research about improving model writing "quality," "alerts" about "improper" outputs, and "marketing materials" about using LLMs for writing. Late in the conferral, Plaintiffs even added "fan fiction" to the list of topics.

It is unclear what Plaintiffs are demanding that OpenAI has not already agreed to produce. Nevertheless, to resolve this dispute, OpenAI will continue conferring with Plaintiffs to understand what, specifically, Plaintiffs need that can be obtained expeditiously. For example, OpenAI could reasonably look for studies relating to customer use of the relevant models as a tool for helping book authors, marketing materials advertising the relevant models' usefulness as a tool for book authors, and similar materials. But, consistent with Plaintiffs' prior representations, document discovery should be tailored and proportional to the needs of the case.

Model Inspection

The parties have agreed on an approach for Plaintiffs' inspection of the in-scope GPT production models. *See* ECF 818. OpenAI opposes Plaintiffs' demand to inspect *draft* versions of GPT models because they are beyond the scope of discovery as defined by Judge Stein (ECF 707 at 3), as well as the other reasons set forth in ECF 819. Because Plaintiffs addressed that dispute in a separate motion, OpenAI refers the Court to its opposition brief. ECF 819.

Expert Report Schedule

While OpenAI is not requesting any change to the current fact discovery schedule, the Court's recent orders may necessitate a change to the schedule for expert discovery. Under the existing schedule, OpenAI's fair use-related expert reports are to be served on the same day that Plaintiffs serve their opening reports on copyright infringement. Because the Court has now permitted Class Plaintiffs to proceed with an output-based claim of infringement and because Plaintiffs have taken the position that they may disclose allegedly infringing outputs for the first time in their expert reports, OpenAI will not know what (or even how many) outputs Plaintiffs allege to be infringing at the time OpenAI submits its fair use-related reports. But because fair use is an affirmative defense to infringement, *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994), OpenAI needs to know what the alleged infringement is ***before*** it submits expert reports to ensure that its affirmative defense squarely meets Plaintiffs' allegations. Because this issue impacts the ordering of expert reports and potentially the presentation of summary judgment motions, OpenAI intends to raise this issue with Judge Stein as soon as practicable.

KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

Sincerely,

KEKER, VAN NEST & PETERS LLP	LATHAM & WATKINS LLP	MORRISON & FOERSTER LLP
/s/ Thomas E. Gorman	*/s/ Allison S. Blanco*	*/s/ Caitlin S. Blythe*
Thomas E. Gorman	Allison S. Blanco	Caitlin S. Blythe