SUSMAN GODFREY L.L.P.
A REGISTERED LIMITED LIABILITY PARTNERSHIP
1000 LOUISIANA
SUITE 5100
HOUSTON, TX 77002-5096
(713) 651-9366
FAX (713) 654-6666
WWW.SUSMANGODFREY.COM

401 UNION STREET
SUITE 3000
SEATTLE, WA 98101-2668
(206) 516-3880

ONE MANHATTAN WEST
NEW YORK, NY 10001-8602
(212) 336-8330

1900 AVENUE OF THE STARS
SUITE 1400
LOS ANGELES, CA 90067
(310) 789-3100

JUSTIN A. NELSON
DIRECT DIAL 713-653-7895

E-MAIL JNELSON@SUSMANGODFREY.COM

November 24, 2025

VIA ECF

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007

RE: *In re OpenAI, Inc. Copyright Infringement Litigation* (No. 1:25-md-03143) This document relates to the following Class Cases: Case No. 1:23-cv-08292, Case No. 1:23-cv-10211, Case No. 1:24-cv-00084, Case No. 1:25-cv-03291, Case No. 1:25-cv-03482, Case No. 1:25-cv-03483.

Class Plaintiffs and OpenAI have made progress on issues related to downloading and output discovery, as reflected in the agreements noted in the parties' most recent correspondence. *See* Ex. 1 (documenting agreements and disputes). However, as set forth below, a number of issues remain where Class Plaintiffs seek the Court's assistance to obtain materials central to their downloading and output claims.

I. Downloading discovery:

A. Personal Azure Containers: At OpenAI, both data—including data downloaded from the internet in the first instance and any resulting datasets—and documents are stored on Microsoft Azure. But within Azure there are both (a) shared "containers," to which a large number of accounts have access, *e.g.*, an account which contained all datasets used in a particular project shared with (at least) all employees on that project, *and* (b) "containers" for an individual's personal files, which includes all the files downloaded along with documents by that particular employee. While Plaintiffs understood that the former were being treated as "non-custodial" sources and thus not subject to search terms, OpenAI did not inform Plaintiffs regarding the latter category or the presence of distinct (i.e. non-duplicative) relevant materials in these "personal containers." Plaintiffs only uncovered the insufficiency of OpenAI's search of Azure in late September and early October, when OpenAI revealed it had withheld files an employee downloaded from LibGen [redacted] *See* Dkt. 659; Dkt. 659-3 at 2; Dkt. 659-4 at 145:2-19 ([redacted] Dkt. 659-13 at 160:20-162:2 ([redacted]).

These personal containers reflect the *specific* files used, downloaded, or worked on by that employee. These sources are, in every sense, "custodial sources" and thus they should have been searched as custodial sources long ago. Indeed, as custodial sources, OpenAI was *required* to run search terms across the materials in these containers under the operative ESI protocol and its predecessors. *See, e.g.,* Dkt. 371, Stipulated ESI Protocol, at ¶ 6 (setting forth the process for search terms for custodial sources). But OpenAI did not do so and refuses to do so now.

Instead, OpenAI states only that it performed a "reasonable investigation" of such sources in accordance with OpenAI's belief that these sources are "non-custodial." Yet, even where OpenAI's self-described "reasonable investigation" uncovered relevant materials downloaded from LibGen [redacted]. *See* Dkt. 659. This calls into question the sufficiency of OpenAI's investigation of Azure generally and the personal containers particularly. Moreover, it demonstrates precisely what Plaintiffs are arguing: these "personal containers" are extremely likely to contain material which OpenAI employees downloaded in the first instance. This material is most relevant to Plaintiffs' downloading allegations, but of course is also relevant to Plaintiffs' allegations generally.

This Court should immediately remedy OpenAI's egregious and until recently undisclosed failure. OpenAI's "excuse"—that it is too burdensome to search these personal repositories using search terms—(a) fails simply by describing it because all parties have an obligation to search relevant places; and (b) doubly fails because Plaintiffs could not have known of the presence or relevance of these materials in these personal containers earlier. While Plaintiffs' raised this issue on October 6—and again on November 1, 4, 5, 7, 13, 14, 19, and 22—OpenAI has failed to

even answer the basic question of whether such containers can be searched by filenames, much less any specific burden from doing so.

In order to accommodate OpenAI's (to date unevidenced) burden concerns, Plaintiffs seek a Court order that OpenAI must custodially search, i.e. using search terms, the personal containers of the following individuals: Ben Mann, Alec Radford, Chris Hallacy, Lillian Weng, Greg Brockman, Leo Gao, Vedant Misra, Ian Sohl, Sam Altman, Jeff Wu, Dario Amodei, and Shantanu Jain. OpenAI should also identify the storage account or other place(s) on Azure [redacted] as it appears this may have been a personal container. To the extent it is a personal container, OpenAI should similarly search that container. For the numerous custodians not captured in the above as well as a set of reseachers Plaintiffs have identified as having likely dealt with material from LibGen or other shadow libraries, *see* Ex. 1 at 20, OpenAI should produce the full listing of directories from those individuals' personal containers, at which point the parties can meet and confer regarding specific directories to investigate and, as necessary, produce from.

Rather than a full search of all custodians' personal containers (which Plaintiffs are entitled to), Plaintiffs compromise focuses the search on 12 individuals (9 of the 32 custodians in this case and 3 additional individuals who Plaintiffs have identified as likely to have downloaded materials from shadow libraries). For the remainder, Plaintiffs' compromise will provide the information necessary for the parties to meaningfully narrow the search to the most relevant materials. The Court should thus order Plaintiffs proposed compromise.

B. Access Settings: Class Plaintiffs seek documents showing which OpenAI personnel had access to the downloaded books material (including any resulting datasets). Such material is plainly relevant, as the extent to which OpenAI made such material available internally—and the access controls over such data—bears directly on OpenAI's fair use defense. *See Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1034 (N.D. Cal. 2025) (rejecting fair use where copyrighted books were retained such that "'hundreds of engineers' could access [the books] to make copies for other uses, and engineers did make other copies," where AI company had "dodged discovery on these points," and finding that the AI company was therefore "not entitled to an order [on fair use] blessing all copying that [the company] has ever made after obtaining the data"); *cf. Authors Guild v. Google, Inc.*, 804 F.3d 202, 227 (2d Cir. 2015) (noting that insufficient access controls which exposed copyrighted books "could expose[] the rights holder to destruction of the value of the copyright" and thus prevent a finding of fair use).

While OpenAI has agreed to search for what it calls "data access documents," *e.g.* documents reflecting what OpenAI's internal policies for access settings regarding its datasets are (and were), OpenAI will not agree to provide such documents or information regarding all periods in question based on OpenAI's belief that some of the IT personnel may have departed.

To be clear, Plaintiffs agree with OpenAI's proposal to search for and produce data access settings. But for any period where OpenAI cannot identify any such document, OpenAI should first search for (1) any server access logs reflecting which individuals actually accessed the data in question, and (2) if neither the data access settings nor the server access logs exist, OpenAI should state as much in a declaration which identifies the individuals who may have knowledge regarding these settings (including former employees).

C. Source Code and Documents Reflecting Modifications: Class Plaintiffs seek the source code and documents reflecting the modifications or filtrations OpenAI performed on books downloaded from Shadow Libraries, books datasets which an OpenAI employee purposefully downloaded due to containing or being suspected of containing books, or any code and documents reflecting OpenAI's efforts to purposefully prune larger datasets of, e.g., scraped materials, down to books. This material is direct evidence of the copies which OpenAI made of Plaintiffs works and of the amount copied from each, both of which are relevant to Plaintiffs' direct infringement claims and OpenAI's fair use defense. Moreover the request is narrowly targeted to books; it does not require OpenAI to produce code used for datasets generally or for non-books materials. The Court should thus order OpenAI to search for and produce such code.

While OpenAI states it has produced the source code related to processing the *training* data for the in-suit models, OpenAI refuses to produce any other code covered by Plaintiffs' request—e.g. code related to processing downloaded data which wasn't included in training sets—due to the purported burdens associated with searching for and producing such code. But this material is highly relevant, OpenAI's burdens are unevidenced, and Plaintiffs' request is narrow. The Court should thus grant Plaintiffs' requested discovery.

D. Copyrighted Books Data and Additional Copies Made of Books Datasets: The sets of copyrighted books downloaded by OpenAI, the number of copies OpenAI made of them, and the use of *each* of those copies are relevant to Plaintiffs' infringement allegations and OpenAI's fair use defense. *See, e.g., Bartz*, 787 F. Supp. 3d at 1026-28 ("The Supreme Court tasks us with looking past the subjective intent of the user to the objective use made of each copy. . . . Like Texaco, Anthropic possessed copies it did not put into use in the laboratory . . . Not every copy was even necessary nor used for training LLMs . . . Our analysis must attend to different uses of different copies, and even to different uses of the same copies."). OpenAI should thus search for and produce such sets, and identify the copies made and uses of the same.

OpenAI's protests that such production is burdensome because it cannot catalogue every set of data purposefully downloaded by an OpenAI employee due to containing copyrighted books or identify every unauthorized copy it made of Class Plaintiffs' works. But in addition to providing no support for its assertion of burden, OpenAI's position is contradicted by the fact that OpenAI appears to have internally identified copies it made of books obtained from, e.g., LibGen as part of its efforts to delete LibGen. *See* Dkt. 846 at 12 (discussing the list of "links to internal OpenAI files related to Library Genesis" in Log No. 18 and ordering production).

The Court should order OpenAI to produce datasets purposefully downloaded due to containing or being suspected of containing copyrighted books and datasets which OpenAI purposefully created via filtering larger datasets of, e.g., scraped materials, down to books.

As to the number of copies made and the uses of these sets, Plaintiffs offered the following compromise which OpenAI rejected and which should be ordered by the Court given the highly relevant nature of this material and OpenAI's failure to substantiate its burden claims: OpenAI shall respond to an interrogatory substantively of the form "(1) including the initial download of the raw data, how many copies did OpenAI make of data downloaded from [shadow library sources identified in Revised Interrogatory Nos. 21-25], (2) identify all such copies and any additional use made of such copies, (3) including the initial download of the raw data, how many copies did OpenAI make of data which comprised the following datasets [list provided for

exemplary purposes: [redacted],[1] (4) identify all such copies and any additional use made of such copies."

To the extent OpenAI made too many copies to count, it should say so; to the extent OpenAI cannot identify all uses, it should say so, indeed Judge Alsup previously faulted another AI company's failure to disclose precisely this information. *See Bartz v. Anthropic*, 2025 WL 2308091 at *1 (N.D. Cal. Aug. 11, 2025) (denying motion for stay and noting that "even now Anthropic has refused to come clean" regarding which works were kept for which uses). OpenAI cannot hide behind burden to avoid saying how many copies it made and what it did with them.

E. Torrenting Material: [redacted] OpenAI downloaded from LibGen was accomplished by OpenAI employee Ben Mann via torrenting—a peer-to-peer file sharing technology that involves downloading users simultaneously sharing the file with third parties. [redacted] Plaintiffs seek targeted documents in order to verify OpenAI's claim and identify any other instances in which OpenAI may have torrented Plaintiffs' works. *See* Ex. 1 at 10 (listing the specific categories). These documents would demonstrate the settings used for OpenAI's torrenting, where the torrents were downloaded, and server logs demonstrating to what extent OpenAI uploaded material while torrenting. Put simply, this material will assist Plaintiffs in determining what books OpenAI downloaded and shared with third parties. *See, e.g., Kadrey v. Meta,* Dkt. 470 (Mar. 7, 2025 N.D. Cal.) (minute order granting plaintiffs' request for torrenting material including an RFP directed at torrenting settings and billing records).

While OpenAI has provided some documents regarding Mr. Mann's torrenting, significant gaps remain and OpenAI has refused to further search for other instances of torrenting or produce documents regarding the same [redacted] *But see* Dkt. 659-13 at 169:13-24 [redacted] But OpenAI's [redacted] underscores the need for investigation given the direct relevance to Plaintiffs' downloading claim and Plaintiffs' claims generally. First, torrents are notoriously used to pirate material—as they were here—and thus identifying other instances of torrenting is exceedingly likely to reveal additional instances in which Plaintiffs' copyrighted works were pirated. Second, any failure by OpenAI to prevent distribution of Plaintiffs' copyrighted works to third parties is categorically not a fair use of the material and must not escape investigation. *See, e.g., Sony Music Ent. Inc. v. Does 1-40*, 326 F. Supp. 2d 556, 565 (S.D.N.Y. 2004) (collecting cases) ("Moreover, the use of P2P systems to download and distribute copyrighted music has been held to constitute copyright infringement."). The Court should thus order (1) OpenAI to search for and produce (or if they have been deleted or cannot be located, confirm as much in writing) the categories of documents Plaintiffs have identified regarding Mr. Mann's torrenting, *see* Ex. 1 at 10 (listing categories) and (2) OpenAI to conduct an investigation regarding any other instances of torrenting data at

[1] These datasets will comprise those which Plaintiffs have so far been able to identify that consist mostly or entirely of books, materials sourced from shadow libraries, materials sourced becaue OpenAI believed they contained books, or materials filtered from webscrapes on the same basis. Plaintiffs expect the productions ordered by the Court's recent opinion regarding waiver will identify additional datasets within these categories which OpenAI should similarly answer. *See* Dkt. 846 at 12 (discussing list of "links to internal OpenAI files related to Library Genesis").

OpenAI (including when an employee's personal computer is used to acquire data later transferred to OpenAI) and, if no other instances are identified, confirm as much in writing. If other instances are identified, OpenAI should produce the same documents (or written confirmation of non-existence) as would be produced as a result of the investigation into Mr. Mann's torrenting of LibGen.

F. LibGen Documents: OpenAI is withholding 859 documents which hit on the search terms regarding LibGen. A subset of those documents—approximately 200 according to OpenAI—are those which OpenAI claimed privilege over but which the Court has now held are either not privileged or waived. *See* Dkt. 846. The remaining ~650 documents, OpenAI claims are "non-responsive."[2] Given core relevance of OpenAI's conduct with respect to LibGen, the Court's order regarding OpenAI's waiver of privilege, and the limited number of apparently withheld documents, OpenAI should immediately produce all 859 documents that hit on the LibGen terms to Plaintiffs.

II. Output Discovery:

A. Output Logs: At this time, Class Plaintiffs are seeking *only the exact same material* which will be made available to News Plaintiffs following the final resolution of the dispute regarding OpenAI's production of ouput logs. Barring any reconsideration, that would consist of the output logs this Court ordered produced at Dkt. 734. This material contains *direct evidence* of Class Plaintiffs' output claims and OpenAI has articulated no unique burden particular to Class Plaintiffs' access to this material. Despite this entire lack of support, OpenAI has refused point blank to produce this material regardless of the outcome of its challenges to the Court's Order requiring production of output logs. *See* Dkts. 742, 746, 752. OpenAI's refusal directly violates the operative protective order in this matter which requires cross-production of all materials which are not printed copies of Training Data or Source Code. *See* Dkt. 367 at ¶ 6, 6.c (cross production obligations with exceptions only for printed copies of source code and training data).

Class Plaintiffs ask is simple: treat Class Plaintiffs' access to logs to support their output claims exactly the same as News Plaintiffs' access to output logs to support their output claims. The Court should order OpenAI to comply with this Court's past (and, to the extent such orders are later modified, future) orders on production of output logs in this MDL.

B. Books Summarization Datasets and Technical Documents: OpenAI [redacted] to develop and improve its models' abilities to summarize books. *See* Ex. 2 at 140:3-141:4; 143:2-14, 150:23-151:10 (excerpts from Ouyang Depo. Tr.). While OpenAI has agreed to search for and produce certain documents regarding these projects, it has refused to collect or produce technical documents or datasets from the same. *See* Ex. 1 at 7. That refusal is unwarranted as (1) the projects used [redacted] and (2) the capacity of OpenAI's models to summarize Plaintiffs' works is directly at issue in Plaintiffs' direct infringement claims and in OpenAI's fair use defense. This is a defined and narrow set of material of core relevance which the Court should order OpenAI to produce.

[2] These would not presently include hits returned on personal containers of OpenAI employees discussed above.

Sincerely,

/s/Justin A. Nelson

Susman Godfrey, LLP
Interim Lead Class Counsel