KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

August 29, 2025

VIA ECF

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl Street
New York, NY 10007-1312

cc: *All Counsel of Record (via ECF)*

Re: ***OpenAI's Response to Class Plaintiffs' August 22, 2025 Letter Regarding Books1&2***
In re OpenAI Inc. Copyright Infringement Litigation, No. 1:25-md-03143-SHS-OTW
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.

Dear Judge Wang:

Class Plaintiffs' filing, a misleading and one-sided six-page "timeline," is most notable for the facts it omits. *First*, no OpenAI model currently available has been trained on books from LibGen. *Second*, the last model trained on such data was GPT-3.5, which was released in 2022 and no longer available on ChatGPT by early 2023. *Third*, OpenAI removed the Books1 and Books2 datasets in July 2022—a full year ***before*** any plaintiff filed suit challenging OpenAI's training on books. *Fourth*, contrary to Plaintiffs' suggestion, OpenAI has recovered the Books1 and Books2 datasets used to train GPT-3 and has produced them to Plaintiffs—and since a subset of those same datasets was used to train GPT-3.5, that data has been recovered and produced too.

But what Plaintiffs' misleading timeline conclusively demonstrates is that Plaintiffs already have the relevant non-privileged facts regarding these datasets and there is no legitimate basis to invade OpenAI's privilege.

The fundamental question in this copyright litigation is whether OpenAI's ***use*** of copyrighted works to train its large language models is fair use. Discovery should be focused on information relevant to that question. The acquisition and use of Plaintiffs' works may be relevant to that question, and Plaintiffs have already had the opportunity to explore non-privileged facts about the Books1 and Books2 datasets to test their allegations. OpenAI has provided 30(b)(6) testimony and other discovery regarding: when the datasets were downloaded, who was involved, and how the data was downloaded, *see* Ex. 1 (7/25/2025 OpenAI 30(b)(6) deposition) at 19:12-20:18, 70:25-72:3, 170:19-171:9; the lawyers and OpenAI employees involved in the decision to remove the datasets, the general subject matter of those discussions, and when and where they took place, *see id.* at 58:1-60:18, 63:16-21; and, when the datasets were removed, by whom, and the technological steps taken for the removal, *see id.* at 78:3-7, 86:13-88:13, 126:18-128:24. OpenAI has also provided testimony regarding its subsequent recovery of the Books1 and Books2 datasets and their production to plaintiffs. *See id.* at 45:13-20, 65:9-70:21, 142:22-143:17. Plaintiffs have the non-privileged facts they need.

What Plaintiffs want now is ***not*** discovery relevant to their copyright infringement claims—such as the contents of the datasets or details about their use in training OpenAI's models. Instead, Plaintiffs seek discovery into the privileged ***reasons*** for ***removal*** of the datasets from OpenAI's systems. But those reasons are, and always have been, privileged because they were based on legal advice. And the reasons for removal are not relevant to whether Plaintiffs' works were infringed. As OpenAI has made clear, there are ***no*** non-privileged reasons for the removal, and it does not intend to rely on the advice of counsel either affirmatively or in rebuttal to Plaintiffs' case-in-chief.

Plaintiffs' repeated attempts to probe the privileged ***reasons*** for ***removal*** is not a legitimate effort to obtain relevant discovery. Only two explanations exist for their persistence. *First*, Plaintiffs may be trying to cast OpenAI's removal of the datasets as improper or nefarious. But these same plaintiffs' lawyers took the exact opposite position in *Bartz v. Anthropic*, arguing that it was wrongful for Anthropic to "hoard" works it was no longer using. *Second*, Plaintiffs may be trying to prop up their willfulness theory. But that theory concerns OpenAI's creation and use of the training datasets, not their later removal by different employees years afterward. In any event, Plaintiffs have all the evidence they need to argue that OpenAI downloaded books from LibGen, that the datasets contained copyrighted works, and that OpenAI removed the datasets from its systems. OpenAI has provided non-privileged discovery on all of those topics. What Plaintiffs cannot do is try to meet their burden by intruding on privileged communications when OpenAI has not drawn the privilege as a sword and has unequivocally disclaimed any reliance on advice of counsel defense.

Ever since Plaintiffs began raising this issue, OpenAI has been consistent: the reasons for the removal of the Books1 and Books2 datasets are privileged. There has been no "flip-flopping" because OpenAI has never attempted to rely on the advice of counsel, has never wielded privilege as a sword and shield, and has never put privileged communications at issue by relying on them to support a claim or defense. Plaintiffs' crime-fraud argument is equally meritless—there has been no crime or fraud, only fair use.[1]

[1] Plaintiffs have abandoned their misguided request that the Court issue a premature evidentiary ruling purporting to bar OpenAI from responding at trial to unidentified and hypothetical future "statements Plaintiffs *may* make" regarding the reasons for the removal. *See* ECF 413 at 4; ECF 479 at 7-8. As OpenAI previously explained, there is no basis for any such ruling and evidentiary rulings are reserved for the trial judge. *See* ECF 428 at 3; ECF 428-1 at 41:16-19.

A. OpenAI has not "flip-flopped" or waived privilege.

There has been no flip-flopping or waiver. OpenAI has never tried to use any privileged communications as evidence of state of mind and is ***not*** relying on the advice of counsel either affirmatively or in rebuttal. *See* ECF 413-11 at 1. Because OpenAI is not "assert[ing] a claim or defense that [it] intends to prove by use of the privileged materials," and "the privileged materials [are not] indispensable to [OpenAI's] claims or defenses," an "at-issue waiver [has not] occur[ed]." *GLD3, LLC v. Albra*, 2024 WL 4471672, at *7 (S.D.N.Y. Oct. 11, 2024). No sword is being drawn; only the shield of privilege is being maintained.

Ever since Plaintiffs first raised this issue in January 2025, OpenAI's position has been consistent: the reasons for the removal of the Books1 and Books2 datasets are privileged. At the January 2025 custodial 30(b)(6) deposition, OpenAI's witness explained that the decision to remove the Books1 and Books2 datasets "was made in consultation with the company attorneys," and he could not "answer further without revealing privileged information." *See* Ex. 2 (1/29/2025 OpenAI 30(b)(6) deposition) at 61:19-62:3.

OpenAI's position remained the same when the issue was briefed and argued at the May 27, 2025 conference. OpenAI stated that its "decision to delete two datasets well before these cases were filed was made in 'consultation with the company attorneys,'" and that "the reasons why are privileged." ECF 390, Case No. 1:23-cv-08292, at 2. In the joint chart submitted in advance of that same discovery conference, OpenAI reiterated that "[t]he reasons for the deletion are privileged." ECF 53-3 at 9.

Plaintiffs try to wring ambiguity out of one snippet of the May 27 hearing transcript. But as the full transcript demonstrates, OpenAI's counsel was drawing a distinction between the fact that the Books1 and Books2 datasets were not being used at the time they were removed, and the privileged reasons for deletion. As OpenAI's counsel explained, allowing discovery into the fact of non-use of the datasets at the time of their removal (which OpenAI has never claimed privilege over) did ***not*** mean that OpenAI was waiving privilege over the ***reasons*** for removal. *See* ECF 413-9 at 70:2–3, 70:9–12 ("just because there may be an element of this issue that is not privileged [*i.e.*, the fact of non-use at the time of removal], doesn't mean that every aspect of the issue is not privileged [*i.e.*, the reasons for the removal].").

After the May 2025 discovery conference, OpenAI took further steps to confirm that it was not putting at issue any privileged communications concerning the reasons for removal of the datasets.

First, OpenAI revised two 2024 letters on unrelated discovery matters that, in passing, stated that the Books1 and Books2 datasets were removed "due to" non-use. Those letters were never intended to disclose the reasons for removal, which OpenAI has always maintained are privileged. Rather, the letters explained that because OpenAI had ceased using the datasets and removed them from its systems long before this litigation, it was still investigating whether it could locate copies of the datasets (they were later successfully located and produced).[2] To

[2] *See* ECF 188-2 at 3 ("[T]he Books1 and Books2 datasets were deleted before any litigation had been filed against OpenAI, and . . . OpenAI has been actively investigating to determine whether

eliminate any doubt about waiver given Plaintiffs' arguments, OpenAI revised the letters in June 2025 to omit the "due to" non-use phrasing.

Second, OpenAI expressly confirmed to Plaintiffs that it would not rely on any non-privileged reason for the removal of the Books1 and Books2 datasets in the litigation. These steps reinforced OpenAI's longstanding and consistent position: the reasons for removal of the datasets are privileged.

Because OpenAI has not waived privilege, there is no basis to compel production of privileged communications. Nor has OpenAI flip-flopped: the reasons for removing the dataset are, and always have been, privileged.

The only flip-fopping has been by plaintiffs' counsel, who have taken directly opposing positions in this case and in *Bartz v. Anthropic*. In *Bartz*, the court suggested that books downloaded from LibGen ***should be removed*** once no longer in use. 2025 WL 1741691, at *16 (N.D. Cal. June 23, 2025). There, Plaintiffs' counsel argued that "hoard[ing]" such unused data was itself infringement. *See* ECF 275, *Bartz v. Anthropic*, Case 3:24-cv-05417-WHA (N.D. Cal., July 28, 2025), at 7. Yet here they contend that removing unused data is itself suspect and wrongful. That contradiction underscores the incoherence of their theory: they cannot plausibly claim that it is unlawful to ***retain*** LibGen data in one case and unlawful to ***remove*** it in another.

B. Plaintiffs do not come close to establishing the crime-fraud exception.

Plaintiffs' "crime-fraud" theory is as inflammatory as it is baseless. Their timeline, misleading from the outset, does not justify piercing OpenAI's privilege. *See* ECF 479 at 8. There has been no crime or fraud—only lawful conduct protected as fair use. Plaintiffs identify no evidence that OpenAI's privileged communications about the reasons for the datasets' removal were made "in furtherance of contemplated or ongoing criminal or fraudulent conduct" and were "*intended* in some way to facilitate or to conceal the criminal activity." *In re Grand Jury Subpoenas Dated Sept. 13, 2023*, 128 F. 4th 127, 141–42 (2d Cir. 2025).

Whether training large language models on copyrighted works is fair use goes to the very heart of this case—it is a merits question to be addressed at summary judgment or trial, not on a discovery motion. Plaintiffs cannot short-circuit that analysis merely by uttering the shibboleth "crime-fraud." To the contrary, courts "should exercise considerable caution" when asked to pierce the privilege at this early stage, particularly where the claim is weak and touches on the merits. *In re Omnicom Grp., Inc. Sec. Litig.*, 233 F.R.D. 400, 407 (S.D.N.Y. 2006).

That caution is particularly warranted here. At least one court to address this issue has rejected Plaintiffs' position outright. In *Kadrey v. Meta Platforms, Inc.*, Judge Chhabria denied plaintiffs'

it can locate any additional copies of the datasets or other documents from which the datasets can be reconstructed.") (as revised); ECF 188-4 at 2 ("We also understand that the use of books1 and books2 for model training was discontinued in late 2021, after the training of GPT-3 and GPT-3.5, and those datasets were then deleted in or around mid-2022. We therefore have not yet located copies of books1 and books2 to make available for inspection, but are actively working on doing so, to the extent that they are still available.") (as revised).

crime-fraud theory from the bench, calling counsel's rhetoric "over the top" and declaring "the crime fraud issue is over." *See* ECF 381-1, Case No. 1:23-cv-08292, at 16:18-18:1. On summary judgment, the court confirmed that Meta's torrenting of books from LibGen is fair use: "Because Meta's ultimate use of the plaintiffs' books was transformative, so too was Meta's downloading of those books" from LibGen. *See Kadrey v. Meta Platforms, Inc.*, 2025 WL 1752484, at *12 (N.D. Cal. June 25, 2025). The same reasoning applies here. If the underlying downloads were lawful, then later removing those datasets cannot possibly be a crime—and communications about the removal cannot further or conceal a nonexistent crime.

Plaintiffs' fallback theory—that consulting counsel about removal a full year ***before*** this litigation was filed somehow constituted "litigation misconduct"—fares no better. OpenAI had no duty to preserve simply because its 2019 PTO submission acknowledged "substantial uncertainty" around fair use in AI training.[3] Legal uncertainty does not trigger a preservation duty; only reasonably foreseeable litigation does. *See In re Keurig Green Mountain Single-Serve Coffee Antitrust Litig.*, 341 F.R.D. 474, 532 (S.D.N.Y. 2022) ("[C]ourts recognize[] that a party's obligation to preserve relevant documents arose when litigation was filed or, at least, ostensibly threatened."). The datasets were removed a full year ***before*** any lawsuit was filed challenging OpenAI's training on books and before OpenAI had even decided to launch ChatGPT. *See* ECF 424-1 (7/31/2025 Kwon Decl.) ¶¶ 7–8. No duty to preserve existed. The involvement of outside counsel does not suggest otherwise. Responsibly consulting with counsel to seek legal advice to assess risk or ensure compliance does not trigger a preservation obligation. *See Cruz v. G-Star Inc.*, 2019 WL 4805765, at *10 (S.D.N.Y. Sept. 30, 2019) (rejecting argument that litigation was likely or that the duty to preserve arose when defendant sought legal advice from outside litigation counsel about the employee who ultimately became a plaintiff in the case).

Plaintiffs' demand for the extraordinary, and disfavored, remedy of piercing the privilege is particularly inappropriate because it is premised on mischaracterizations of evidence and omissions of salient facts. For example, Plaintiffs misleadingly claim that OpenAI has "never recovered" the "datasets used to train GPT-3.5." ECF 479 at 4. But the versions of the Books1 and Books2 datasets that were used to train GPT-3.5 are ***subsets*** of the recovered and produced versions that were used to train GPT-3. *See* Ex. 2 at 58:21-59:2, 59:11-13 ("It's a subset of Books1 and Books2. . . . [t]here was a subset of Books1 and Books2 used for the training of [] GPT 3.5."). In other words, Plaintiffs have access to all the Books1 and Books2 datasets used to train both GPT-3 and GPT-3.5.

As another example, Plaintiffs assert that "OpenAI delete[d] . . . the original files downloaded by Mann and Radford" in 2022. *See* ECF 479 at 3. That is incorrect. OpenAI has produced Radford's raw files, and Mann's raw files were ***not*** "deleted" in 2022.[4] Plaintiffs' assertion that

[3] Plaintiffs' position on so-called "uncertainty" is based on the improper assumption that the fair use question must go their way. In fact, OpenAI's 2019 submission to the PTO explained why training was fair use. The reference to "uncertainty" was tied to the fact the technology was new and courts had not yet addressed it.

[4] As OpenAI previously explained to Plaintiffs, the evidence to date suggests that Mann's raw download was either not retained at the time, or it was not migrated to OpenAI's new Azure

"OpenAI delete[d] . . . any references in source code to [LibGen1 and LibGen2]" is also incorrect. *See id*. There was no deletion of source code. OpenAI's 30(b)(6) witness testified that, although OpenAI [redacted] so that any [redacted] the prior versions of the source code preceding those modifications have been retained and produced in this action. Ex. 1 at 86:19-87:12. And finally, OpenAI's 30(b)(6) witness did not testify that Morrison & Foerster advised OpenAI to delete LibGen—the cited testimony does not say that. *See* ECF 479 at 6. Rather, Morrison & Foerster "gave advice" in connection with the issue. Ex. 1 at 91:7-15.

Indeed, removal of the Books1 and Books2 datasets was not only lawful but, according to one federal judge, a prudent course of action. As Judge Alsup explained in *Bartz*, the problem was that Anthropic retained books obtained from LibGen "even after deciding it would not make further copies from them for training," which "indicat[ed] there were other further uses." 2025 WL 1741691, at *16. By contrast, here, OpenAI ***removed*** the Books1 and Books2 datasets after they were no longer being used and well ***before*** these lawsuits were filed—that is certainly not evidence of any crime or fraud.

Respectfully,

KEKER, VAN NEST & PETERS LLP[5]	LATHAM & WATKINS LLP	MORRISON & FOERSTER LLP
/s/ R. James Slaughter	*/s/ Margaret Graham*	*/s/ Caitlin Sinclaire Blythe*

storage system in 2020.

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