September 19, 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 (S.D.N.Y.). This Document Relates To: 23-cv-11195

Dear Magistrate Judge Wang:

Pursuant to the sampling protocol entered by the Court on August 19, 2025 (MDL ECF 463) (the "Protocol"), OpenAI Defendants and News Plaintiffs write to present each side's positions regarding the sampling exercise.

OpenAI's Position: The hit counts OpenAI provided confirm what OpenAI has said all along: "[i]n the real world, people do not use ChatGPT" in the way that News Plaintiffs claim they do. *See* NYT ECF 52. The frequency with which users ask ChatGPT about topics that are relevant to News Plaintiffs' claims is vanishingly low. User mentions of "wirecutter" are roughly one-in-a-million. "[N]ew york daily news" is even less frequent. So too for the 89 so-called "pink slime" websites. Less than [redacted] percent of prompts include references to the New York Times, which (in all likelihood) would include references that have nothing to do with these claims (*e.g.*, a question about Times's stock price or history). News Plaintiffs attempt to skew these numbers by lumping together 50 terms and reporting aggregated hit counts. But the terms they lump together include generic terms like "twin cities," "intercept," "tribune," "daily news," and "cir," which no doubt result in thousands of false positives and account for roughly two-thirds of the aggregated hit counts they report. That is precisely why the Protocol precludes the use of these hit counts for merits purposes. Protocol § 5. The way to draw conclusions about the content of ChatGPT conversations is by sampling the more than 40 billion conversations that OpenAI has retained and made available for merits analysis.

More to the point: News Plaintiffs fail to address the singular purpose of this exercise: to determine whether OpenAI's policy of honoring users' privacy decisions has resulted in a "bias in what will be [] sampled" for merits analysis. 05/27/2025 Hr'g Tr. at 4:13–20. News Plaintiffs previously claimed that a bias exists because the data available for merits sampling does not contain conversations generated via ChatGPT's Temporary Chat feature ("**Temporary Chats**") and conversations that users manually delete via the ChatGPT interface ("**Deleted Chats**"), which OpenAI retains only for 30 days. But the hit counts prove that the exclusion of Temporary and Deleted Chats will have no discernible effect. That eliminates any claim that News Plaintiffs have suffered prejudice under Rule 37(e), *see* ECF 217 (purpose of protocol is to "aid in the Court's Rule 37(e) analysis"), along with the justification for the Preservation Order, *see* 05/27/2025 Hr'g Tr. at 4:13–20 (Court explaining that the Preservation Order was issued due to concern about sampling bias). The Court should vacate the Preservation Order in its entirety.

As referenced above, the question before the Court is whether a statistically significant difference exists between **(i)** the data available for sampling—which *excludes* Temporary and Deleted Chats—and **(ii)** the data that would have been available for sampling had OpenAI retained Temporary and Deleted Chats indefinitely. The Protocol provides for two data samples that correspond to those two populations: **(i)** a "**Retained Sample**" consisting of logs drawn from a population of all conversation data *other than* Temporary and Deleted Chats, *see* Protocol § 1(c)(i); and **(ii)** an "**All-In Sample**" consisting of logs drawn from a population representing all chats, *including* Temporary and Deleted Chats, *see id.* § 1(c)(iv). If News Plaintiffs are correct that honoring users' privacy choices will cause prejudicial sampling bias, then that bias would have shown up in the form of higher hit counts for the relevant terms in the All-In Sample (which includes Temporary and Deleted Chats) than in the Retained Sample (which does not).

The hit counts show the opposite. For example, the hit counts for "nytimes" in the Retained and All-In Samples are effectively the same: the term appears in [redacted] prompts and [redacted] outputs in the Retained Sample, and [redacted] prompts and [redacted] outputs in the All-In Sample. (Both samples include 5 million logs in total.) Declaration of Maxime Cohen, Ex. B2 ("Cohen Decl."). These are not statistically significant differences. *Id.*. In fact, OpenAI's expert analyzed each of the hit counts in the Retained Sample and the All-In Sample and found that out of the ***288 data points***, only ***six*** are statistically significant. *Id.*, Ex. B2, B3. Three of those are statistically significant ***in the opposite direction***—meaning that the terms that Plaintiffs chose are actually more frequent in the Retained Sample (which corresponds to the data population available for sampling). *Id.* ¶¶ 22–25. And the remaining three, which relate solely to differences in hits on outputs (not prompts), do not indicate a material sampling bias upon further analysis (*e.g.*, when analyzing "paywall" and "pay wall" together). *Id.* ¶¶ 27–31.

News Plaintiffs do not even attempt to engage in a proper statistical analysis. Instead, they focus on a separate sample that they insisted on including in the Protocol (over OpenAI's objection, *see* ECF 215-3 at 12) composed entirely of Temporary Chats. But as News Plaintiffs well know, Temporary Chats comprise ***less than*** [redacted] of all ChatGPT conversations. ECF 27-1 at 3. So even if ChatGPT users are slightly more likely to mention the word "paywall" when using the Temporary Chat feature, that does not suggest that there will be "bias in what will be [] sampled" for merits purposes, particularly when the hit counts show the Retained and All-In Samples are materially identical. Cohen Decl. ¶¶ 6–12.[1] News Plaintiffs' aggregated hit counts do not help them either: even setting aside the fact that their table is highly misleading, *see supra*, these hit counts do not even purport to show prejudicial sampling bias. To the contrary, News Plaintiffs' table suggests that their publication names are *more frequent* in the data available for sampling than in Temporary and Deleted Chats.

Finally, News Plaintiffs demand that the Preservation Order remain in effect for "output log data that hits on the search terms and classifiers used as part of this sampling exercise." But the parties are well underway on a process to retrieve a representative sample of conversation log data for

[1] To reprise the Court's analogy: if you want to evaluate whether the removal of a subset of gumballs from a gumball machine will bias a sample of gumballs that the parties plan to analyze, 05/27/2025 Hr'g Tr. at 5:18–6:8, that requires an analysis of **(1)** what gumballs were in the machine before the removal (*i.e.*, All-In Sample) and **(2)** what gumballs were in the machine after the removal (*i.e.*, Retained Sample). If those two distributions are effectively the same, there is no sampling bias. Looking only at the gumballs that were removed does not call that into doubt, particularly if (*e.g.*) only 5 gumballs were removed from a machine that previously contained 10,000.

merits analysis. News Plaintiffs do not explain what they hope to do with the logs they want preserved that they cannot do with that representative sample. In any case, this Court emphasized at the May hearing that the preserved data will not be "provide[d] to anybody" and even the Court will not be "looking at it." 05/27/2025 Hr'g Tr. at 5:15–17. The Court should deny that request and vacate the Preservation Order in its entirety.

News Plaintiffs' Position: The parties and the Court are here because OpenAI chose to destroy over 8 billion ChatGPT output logs after The Times filed suit in December 2023, and up until the Court's May 13, 2025 Preservation Order. *MDL* Dkt. 33. That Order was premised in part on this Court's hypothesis that users who were using ChatGPT to evade publishers' paywalls were also more likely to ask OpenAI to delete their chats. Dkt.43-2 at 38:14-20. The hit report attached as Exhibit A confirms this Court's hypothesis. Specifically, that hit report shows that users were in fact far more likely to engage in ChatGPT prompts and outputs classified as news-related (and potentially paywall evading or otherwise interfering with publishers' ability to monetize their content) as part of a "temporary chat" (that is, one where the user had requested that her queries and outputs not be retained) than chats that were ordinarily retained. Those "temporary chats" are among the records that OpenAI had, until the Preservation Order, been destroying. The hit report also suggests the presence of significant (millions of) instances of potential output-based infringement in the output log data OpenAI has been ordered to preserve. News Plaintiffs therefore respectfully submit that OpenAI should continue to preserve output log data that hits on the news-related classifiers and keywords used for this exercise at least until the conclusion of the litigation.

The hit reports show that temporary chats have statistically significant higher proportions of conversations relating to news content as compared to the ordinarily retained chats:

(1) Conversations classified as relating to information about public people, information about current events and news, and purchasable products occur **57.3%**, **46.5%**, and **18.9%** more frequently, respectively, in temporary chats as compared to ordinarily retained chats;

(2) Queries and outputs including the "News" search term[2] appear **20%** and **11%** more frequently, respectively, in temporary chats as compared to ordinarily retained chats; and

(3) Queries and outputs including "pay wall" and "paywall" appear **33.8%** and **70.8%** more frequently, respectively, in temporary chats as compared to ordinarily retained chats.

See Declaration of Tom Goldstein, ¶¶ 6, 8, 10. These statistics support the proposition that users looking to evade paywalls and get news content (including News Plaintiffs' content) for free were, and are, more likely to do so using the temporary chat feature—knowing that OpenAI will delete their conversations.

The hit reports also include an astonishing number of hits on News Plaintiffs' names in both the prompts and the outputs across the different populations of ChatGPT conversations—notwithstanding that OpenAI had implemented plaintiff-specific blocks over a year ago that likely deterred many users from continuing to search for News Plaintiffs' content. The table below includes aggregated hits on News Plaintiff's names and variations thereof—The Times, *Daily News*, and the Center for Investigative Reporting—for both prompts and outputs:

[2] The "News" search term is "news AND (article* OR headline* OR breaking OR current OR summar* OR today OR yesterday OR week OR happen* OR occur* OR status OR update*)".

	Ordinarily Retained Data	Temporary Chat	User-Initiated Deleted Chat
Prompts	[redacted]	[redacted]	[redacted]
Outputs	[redacted]	[redacted]	[redacted]

The prompts that include a News Plaintiffs' name is about [redacted]% and [redacted]% of the 5 million samples for the temporary chat and user-initiated deleted populations, respectively—namely, but for the Preservation Order, OpenAI would have deleted the approximately [redacted]% of user prompts to ChatGPT using News Plaintiffs' names in April-May 2025, and the associated output. Because there are, according to Sam Altman, 330 million daily US-based chats,[3] these numbers extrapolate out to approximately [redacted] user prompts on a daily basis that mention News Plaintiffs' names, or nearly 4 million prompts per month—each of which is potentially the basis for a statutory damages claim.[4] These numbers expand roughly 3x when considering the number of times a News Plaintiffs' name appears in the output alone.

What is missing from this exercise (due to OpenAI's destruction of past output log data) is how user behavior has changed over time in view of OpenAI's blocks on News Plaintiffs' content. For example, many users likely gave up searching for News Plaintiffs' content by name after repeatedly encountering (for more than a year) OpenAI's blocks that, according to OpenAI, prevent News Plaintiffs' content from appearing in output. Accordingly, the general "news" terms and the news-related classifiers are more probative in ascertaining whether there are differences between the populations of data with respect to paywall-avoiding activities, and the prevalence of hits on News Plaintiffs' names (and potentially infringement) from this exercise are almost certainly underrepresented as compared to historical populations. Moreover, these hit reports do not reflect infringement due to retrieval augmented generation because such data is not included in the data OpenAI preserved.

In view of these findings, News Plaintiffs respectfully request that the Court maintain the Preservation Order, but limit the scope, on a moving forward basis, to all ChatGPT and API output log data that hits on the search terms and classifiers used as part of this sampling exercise.[5]

Respectfully submitted,

KEKER, VAN NEST & PETERS LLP
/s/ Edward A. Bayley
Edward A. Bayley
On behalf of OpenAI Defendants

ROTHWELL, FIGG, ERNST & MANBECK, P.C.
/s/ Jennifer B. Maisel
Jennifer B. Maisel
On behalf of News Plaintiffs

[3] https://www.axios.com/2025/07/21/sam-altman-openai-trump-dc-fed

[4] Using OpenAI's assertion that "[l]ess than [redacted] percent of prompts include references to the New York Times," that means that [redacted] user prompts in the U.S. each day mention The New York Times. With respect to the "pink slime" websites, [redacted] prompts of the 20 million user prompts that OpenAI sampled correlates to [redacted] user prompts in the data OpenAI has retained. And these are only the prompts input by users. OpenAI has deleted the data that would show ChatGPT retrieving News Plaintiffs' content from such sites.

[5] After the parties exchanged their positions on Wednesday pursuant to the parties' agreed-upon schedule, OpenAI completely rewrote its submission.