**VIA ECF**
Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York

**RE: *In re: OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143 (SHS) (OTW); this document relates to Nos. 1:23-cv-11195, 1:24-cv-01515, 1:24-cv-03285 1:24-cv-04872, 1:25-cv-00501**

Dear Judge Wang:

We write on behalf of the New York Times, the *Daily News* Plaintiffs, The Intercept, the Center for Investigative Reporting, and the *Ziff Davis* Plaintiffs (collectively, "News Plaintiffs") seeking an order compelling OpenAI to produce the original copies of News Plaintiffs' articles that OpenAI obtained from a third party, the Common Crawl Foundation ("Common Crawl"). News Plaintiffs seek to confirm that when OpenAI obtained their articles from Common Crawl, OpenAI also obtained the corresponding copyright management information ("CMI") News Plaintiffs conveyed with those articles. This fact is relevant to News Plaintiffs' CMI removal claims under 17 U.S.C. § 1202(b)(1): to prove that OpenAI removed CMI from News Plaintiffs' articles, News Plaintiffs intend to establish that OpenAI obtained the CMI in the first place and that the CMI had not previously been removed by another entity, such as Common Crawl itself.

There is one natural way for News Plaintiffs to prove this fact: for OpenAI to produce copies of News Plaintiffs' websites in the original form that it obtained them from Common Crawl. If News Plaintiffs' CMI is present on those copies but absent from the copies OpenAI used to train the GPT models—the latter of which is the subject of other discovery News Plaintiffs have conducted—then OpenAI must have removed the CMI. News Plaintiffs served RFPs on May 13, 2025 for copies of their websites in the form OpenAI obtained them. *See, e.g.*, Exs. 1, 2 (Intercept RFP No 12; CIR RFP No. 67).[1] Despite four months of meet and confer efforts, OpenAI has refused to produce those copies, and the Court should compel it to do so.

## I. OpenAI obtained copies of News Plaintiffs' websites from Common Crawl.

Some background on Common Crawl is necessary to understand the dispute. Common Crawl scrapes data from individual websites and publishes its crawls online.[2] Each crawl includes all the websites that Common Crawl scraped in a given scraping cycle. Each crawl is colossal in size because it includes large swaths of the internet.

Common Crawl publishes its crawls at regular intervals; for instance, it has released nine crawls thus far in 2025.[3] But Common Crawl does not merely scrape websites that have gone live

[1] The New York Times, the *Daily News* Plaintiffs, and the *Ziff Davis* Plaintiffs have served similar RFPs. *See* Ex. 3 at 7, 8, 10 (NYT RFP Nos. 75, 77, 81, 93); Ex. 4 at 10, 11 (Daily News RFP Nos. 96, 99, 100, 107); Ex. 5 at 9-10 (Ziff Davis RFP No. 15); Ex. 6 at 9-10 (Ziff Davis RFP No. 32).
[2] News Plaintiffs use the verb "scrape" to denote Common Crawl's act of downloading website content, and the noun "crawl" to denote a set of data that Common Crawl publishes online.
[3] https://commoncrawl.org/get-started.

since the last crawl. It also scrapes websites that have long existed, and, significantly, websites that have been scraped before. Thus, a given news article published in 2018 could have been scraped multiple times per year since its publication, and could be available on many crawls published since then. And because websites can change over time, there may be differences between the website as it was scraped in 2018 and the same website as it was scraped in 2025.

Downloading websites from Common Crawl is extremely data- and resource-intensive. A single crawl contains billions of webpages and hundreds of terabytes of uncompressed data.[4] Downloading and processing even a single crawl, let alone the dozens of crawls currently online, is therefore not practical for news organizations like News Plaintiffs. Common Crawl has an alternative option to search for and download unique URLs, which News Plaintiffs have used to examine a small sample of copies of their websites. But it is unclear the extent to which this can be scaled up given the large amount of data at issue and the computational resources this task would require.

OpenAI has admitted that it used data from Common Crawl data to train its models. Ex. 7 at 14 (Response to Intercept RFA No. 38). It has also admitted that when it downloaded a given crawl, it took whatever CMI the crawl contained. Ex. 8 at 6-10 (Responses to Intercept RFA Nos. 53-56).[5] As News Plaintiffs understand it, moreover, OpenAI downloaded many crawls to populate its training sets, and thus potentially many copies of the same articles. OpenAI has confirmed that it should be feasible to identify which crawls correspond to which training sets. Ex. 9 at 3. But it has not identified those crawls or said when it will do so.

## II. The Court should compel OpenAI to produce those copies.

News Plaintiffs served RFPs seeking copies of their websites in the form OpenAI obtained them. *See, e.g.*, Exs. 1, 2. OpenAI does not appear to dispute the relevance of these documents. And they clearly are relevant to News Plaintiffs' DMCA claims and to the *Ziff Davis* Plaintiffs' section 1202(b)(3) claim. News Plaintiffs have alleged that OpenAI removed their CMI, and for OpenAI to have removed their CMI, it must have accessed the CMI. That is what News Plaintiffs seek to prove through these RFPs.

OpenAI has instead resisted producing the copies because it is theoretically possible for News Plaintiffs to download them from Common Crawl once OpenAI identifies the relevant crawls—which it has neither done nor committed to doing by a date certain. *See* Ex. 9 at 1 ("If you are asking us to produce the WARC files to you, our response is this data is publicly available.").[6]

---

[4] *See, e.g.*, https://commoncrawl.org/blog/september-2021-crawl-archive-now-available. Here Common Crawl describes the crawl's size in "TiB," or "tebibytes," which is slightly larger than a terabyte.

[5] News Plaintiffs still need copies of their websites from the crawls because, while OpenAI has admitted that it took whatever CMI Common Crawl published, it has refused to admit (or deny) that Common Crawl published CMI in connection with News Plaintiffs' asserted works. News Plaintiffs do not presently challenge these RFA responses but reserve all rights to do so.

[6] "WARC" refers to the format of files downloadable from Common Crawl.

As an initial matter, OpenAI has offered no basis to assume that Common Crawl, a third party, has made no changes to the crawls in the years since OpenAI downloaded them to train its GPT models. Indeed, there are affirmative indications that the data has been modified.[7]

In any event, even taken on its own terms, OpenAI's objection is baseless. Rule 26 provides for discovery of documents that are relevant and proportional to the needs of the case. This includes documents that the moving party could theoretically obtain from another source. *See, e.g.*, *Komatsu v. City of New York*, No. 18-cv-3698, 2020 WL 14051597, at *2 (S.D.N.Y. Dec. 18, 2020) ("An objection that the information sought in an interrogatory or request for production is equally available to the requesting party is insufficient to resist a discovery request."); *Au New Haven, LLC v. YKK* Corp., No. 15-cv-3411, 2023 WL 2612204, at *10 (S.D.N.Y. Mar. 23, 2023) ("There is no carve out from the discovery rules that permits a party to fail to produce responsive information merely because the information has been filed publicly."); *Duran v. Cnty. of Riverside*, No. 23-cv-00106, 2024 WL 3468758, at *3 (C.D. Cal. July 3, 2024) ("An objection that documents are equally available to the requesting party is inappropriate") (quoting *Miller v. White*, 2020 WL 4001624, at *6 (C.D. Cal. May 27, 2020)); *Charter Pracs. Int'l, LLC v. Robb*, No. 12-cv-1768, 2014 WL 273855, at *2 (D. Conn. Jan. 23, 2014) (overruling objection that requested documents are "available online" and collecting cases). Courts have made exceptions in "unusual circumstances," such as when producing the document would impair the producing party's own access to it. *Lifeguard Licensing Corp. v. Kozak*, No. 15-cv-8459, 2016 WL 4733157, at *4 n.2 (S.D.N.Y. Sept. 9, 2016). But there is no general carve-out as OpenAI suggests.

Requiring News Plaintiffs to obtain the relevant copies from the crawls, moreover, would make no sense in the context of this case. Due to their size, News Plaintiffs cannot feasibly download the crawls themselves; OpenAI, however, has already done that and can simply produce the relevant portions corresponding to News Plaintiffs' websites. And while it may be theoretically possible for News Plaintiffs to isolate specific URLs for download, it is doubtful that this is practical at the scale that would be required in this case—a scale caused entirely by OpenAI's decision to train its models on a massive number of News Plaintiffs' articles without permission.

In addition, given the nature of the data, one party will likely need to produce it to the other regardless. Because Common Crawl data consists of dense computer code, because a given website may differ across crawls, and due to the sheer number of copies involved, there is only one practical way for the parties to ensure, at summary judgment and trial, that they are referring to the same documents: the documents must be Bates-stamped and produced. And it is surely far more burdensome for News Plaintiffs to obtain ***and*** produce the relevant parts of the crawls than it is for OpenAI to simply produce the parts of the crawls that it already obtained. The Court should order it do so.

[7] https://www.wired.com/story/the-fight-against-ai-comes-to-a-foundational-data-set/.

September 9, 2025

Respectfully submitted,

*/s/ Stephen Stich Match*
Stephen Stich Match
Loevy & Loevy

*/s/ Davida Brook*
Davida Brook
Susman Godfrey L.L.P.

*/s/ Steven Lieberman*
Rothwell, Figg, Ernst & Manbeck P.C.

*/s/ Lacy H. Koonce, III*
Lacy H. "Lance" Koonce
Klaris Law PLLC

cc: All Counsel of Record (via ECF)