KEKER VAN NEST & PETERS LATHAM & WATKINS LLP MORRISON FOERSTER

September 12, 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 News Plaintiffs' motion to compel copies of Common Crawl's public data [Dkt. 561]**; *In re OpenAI 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-04315

Dear Judge Wang:

News Plaintiffs' motion to compel copies of Common Crawl's public data disregards this Court's meet-and-confer requirement and impermissibly tries to force OpenAI to assign its employees or experts to conduct technical analysis that News Plaintiffs can do themselves. Dkt. 561 ("Mot.") at 3. *First*, the Motion violates this Court's Individual Practices and disregards the Court's repeated admonitions to first meet and confer in good faith. The NY Times, Daily News, and Ziff Davis plaintiffs did not meet and confer ***at all***. No email, no call, nothing. This gross failure to meet and confer should not be excused. *Second*, if News Plaintiffs seek copies of data in a form that currently exists, OpenAI's files are the same ones that News Plaintiffs can download publicly from Common Crawl.[1] *Third*, News Plaintiffs may be asking the Court for something they never conferred on. Instead of requesting documents in the form that OpenAI keeps them, News Plaintiffs want to dragoon OpenAI's engineers or consultants into conducting unspecified technical analysis of Common Crawl's public data to locate News Plaintiffs' asserted works, process them, and export new data. News Plaintiffs have ***many*** technical consultants, and nothing in Rule 26 requires OpenAI to do News Plaintiffs' analysis for them.

[1] The Common Crawl Foundation is a 501(c)(3) non-profit that crawls the open web and makes that data freely available to academics, researchers, and entrepreneurs. *See* https://commoncrawl.org/mission

I. Background

This Motion is the ***first time*** that the NY Times, Daily News, and Ziff Davis plaintiffs have cited any RFPs purportedly requesting asserted works that OpenAI obtained from Common Crawl. *See* Mot. at 1 n.2. On their face, the cited RFPs do not seek crawl data but appear to request other materials, such as documents related to ***OpenAI's*** (not Common Crawl's) collection of data from the internet, *see* Dkt. 561-3 at 9 (RFP 81), or documents that show OpenAI's methods of processing data (not the data itself), *see* Dkt. 561-5 at 9-10 (RFP 15). If the NY Times, Daily News, and Ziff Davis plaintiffs disagree, they have never met and conferred with OpenAI to request the production of "relevant parts" of Common Crawl data. Mot. at 3.

As for Intercept and CIR, both plaintiffs propounded an RFP that sought copies of their asserted works "in the form in which [OpenAI] obtained" them. Dkt. 561-1 at 1; Dkt. 561-2 at 2. According to Intercept and CIR, the "original form" in which OpenAI obtained the asserted works will show whether any copyright management information (CMI) was later removed in OpenAI's training datasets. Mot. at 1. OpenAI has already confirmed that it used data from Common Crawl's public data to train its models and that it received whatever CMI is in such data. Mot. at 2. And News Plaintiffs know how to download Common Crawl's data. *See id*. If the Common Crawl Foundation made News Plaintiffs' website data available publicly, News Plaintiffs can access it via Common Crawl's Web ARChive (WARC) files. *See* Dkt. 561-9 at 1. In total, these WARC files contain six pebibytes (or 6.7554 million gigabytes) of data.

Even though Intercept and CIR's RFPs asked for crawl data "in the form in which [OpenAI] obtained" them—*i.e.*, public WARC files—they now appear to pivot to something ***never discussed***. Now they ask the Court to order OpenAI to "produce the ***parts***" of the public WARC files (apparently meaning the asserted works). Mot. at 3 (emphasis added). This would require OpenAI to forensically analyze 6.75 million gigabytes of data. News Plaintiffs admit they could do this analysis themselves but complain that it requires "computational resources." Mot. at 2. Because Intercept and CIR raised this request for the first time in its Motion, the parties have never discussed what methodology could analyze public WARC files to extract whatever portions News Plaintiffs request.

II. Argument

The Court should deny the motion for multiple independent reasons.

First, the NY Times, Daily News, and Ziff Davis plaintiffs failed to meet and confer with OpenAI about the production of ***any*** data from Common Crawl's public data—brazenly ignoring the Court's repeated admonitions that the parties ***must*** meet and confer in good faith ***before*** filing. *E.g.*, 5/27/25 Hr'g Tr. Dkt. 109 at 64:12. Their rush to file this motion means that the parties have ***never*** discussed important issues, such as the specifics of the analysis that News Plaintiffs expect OpenAI to perform and what work product might be generated. Given how

often the Court has admonished the parties to meet and confer, their request for relief should be denied with prejudice to discourage premature and burdensome motion practice.

Second, if Intercept and CIR seek copies of the public WARC files, OpenAI is not required to produce materials that are equally available to them. "It is well established that discovery need not be required of documents of public record which are equally accessible to all parties." *S.E.C. v. Samuel H. Sloan & Co.*, 369 F. Supp. 994, 995 (S.D.N.Y. 1973); *Hart v. Suffolk Cnty.*, 2023 WL 5720075, at *7 (E.D.N.Y. Sept. 5, 2023) (same); *Baum v. Vill. of Chittenango*, 218 F.R.D. 36, 40 (N.D.N.Y. 2003) ("[C]ompelling discovery from another is unnecessary when the documents sought are equally accessible to all."). There is no dispute that (1) Intercept and CIR's RFPs seek "articles . . . in the form in which [OpenAI] obtained those articles"; and (2) OpenAI obtained articles from Common Crawl in the form of a WARC file. Dkt. 561-1 at 1; Dkt. 561-1 at 2; Dkt. 561-9 at 1. Because the ***same*** WARC files are publicly and freely available to Intercept and CIR, they have what they need under the plain terms of the RFPs.

Third, if Intercept and CIR now ask OpenAI to analyze and create new documents excised from publicly available Common Crawl data, that request is also improper. Rule 34 does not require OpenAI to conduct a forensic analysis of 6.75 million gigabytes of data, filter it through a methodology that the parties never discussed, and generate files that are not used in the ordinary course of business and do not currently exist.

Hahn v. Massage Envy Franchising, LLC is directly on point. 2014 WL 12899290 (S.D. Cal. July 24, 2014). There, the plaintiff sought to compel the defendant to "produce relevant excerpts of its [software] database," *id.* at *1, even though the defendant previously "produced the entirety of its database in the format the database is ordinarily maintained." *Id.* at *2. The *Hahn* plaintiff requested that the defendant "run queries" on the data and "convert the resulting data" into his preferred format. *Id.* at *5. The Court denied the plaintiff's motion to compel, holding that "Rule 34 does not require . . . [the] Defendant to perform the laboring oar and run numerous queries and export the results of those queries out of the form it is ordinarily maintained . . . and into an entirely different format." *Id.* at *9. That is especially true here when News Plaintiffs have not even proposed ***how*** the asserted works can be identified and isolated from the WARC files.

Finally, Intercept and CIR's complaint about burden only underscores that they are trying to outsource expert work to OpenAI's engineers and consultants. If Intercept and CIR believe that a technical analysis of this data is necessary to prove their DMCA claims, that examination must be done by their own experts. OpenAI "should not have to conduct a review of records . . . to which [News Plaintiffs] [have] access and make determinations that [News Plaintiffs] can do for [themselves]." *Espinal v. Coughlin III*, 2000 WL 245879, at *1 (S.D.N.Y. Mar. 3, 2000). Indeed, "parties are generally responsible for their own costs." *Baum*, 218 F.R.D. at 40-41. News Plaintiffs cannot credibly argue that they lack the resources to analyze this data. The law firm representing Intercept and CIR employs their own data scientist. The NY Times and Daily News plaintiffs have flown ***at least*** six consultants across the country to inspect

OpenAI's data, and they have collectively spent over 100 business days analyzing OpenAI's data and source code. And the NY Times alone spends millions per quarter on this case, yet still retains an operating profit of $133 million per quarter.[2]

If News Plaintiffs want to extract data from publicly available WARC files, they can do it themselves. Nothing in Rule 26 allows Intercept and CIR to deputize OpenAI's employees or consultants and divert them away from their work to do News Plaintiffs' own expert analysis for them.

Respectfully,

KEKER, VAN NEST & PETERS LLP[3]	LATHAM & WATKINS LLP	MORRISON & FOERSTER LLP
/s/ Thomas E. Gorman	*/s/ Allison S. Blanco*	*/s/ Rose S. Lee*

[2] *See* https://nytco-assets.nytimes.com/2025/08/Q2-2025-Earnings-Release.pdf

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