orrick faegre drinker

November 21, 2025

**VIA ECF**

Honorable Ona T. Wang
United States District Court
500 Pearl Street
New York, NY 10007

Re: *In re OpenAI, Inc., Copyright Infringement Litig.*, 1:25-md-3143-SHS-OTW
This Document Relates To: The Class Cases
Microsoft's Response to Class Plaintiffs' Motion to Compel Inspection of the Bing Index (ECF 777)

Dear Magistrate Judge Wang:

Class Plaintiffs' operative complaint does not mention the Bing Index in any way, shape, or form. This is not surprising given that the case they originally brought was about training of large language models and there is no dispute that the Bing Index was **never used to train any large language model in this case**.[1] And yet, they continue their ever-meandering frolic by bothering the Court with their latest motion. Class Plaintiffs do so even after Microsoft proposed a set of reasonable steps that would address the proportionality concerns addressed by the Court and moot any remaining dispute. Class Plaintiffs' motion does not even address the substance of this proposal, other than a passing reference in a footnote. ECF 777 at 1 n.1.

Microsoft's written proposal (attached in full as **Exhibit A**) pledged, for each of the purported "notorious piracy websites" identified in Class Plaintiffs' motion[2] (less WebText, Internet Archive, and Reddit), to make the following information available from the Bing Index through querying Microsoft's internal Bingdex tool: (1) total number of URLs indexed, per domain; (2) the segregated number of URLs indexed and number of URLs robots disallowed from indexing, per domain; and (3) HTML contents for the URLs associated with each of the domains. In light of what the Court has already found to be the "limited relevance of today's Bing Index," and given that "the parties do not appear to dispute that the near entirety of the Bing Index is irrelevant to this litigation" (ECF 712 at 9), this narrowed scope still amounts to a significant burden on Microsoft with respect to review and production. Requiring Microsoft to produce or make available anything more from the current Bing Index would be both unduly burdensome and wildly disproportionate to the needs of the case, especially when weighing—as required under Rule 26(b)(1)—whether the burden or expense of the proposed discovery outweighs its likely benefit.

This is exactly the result that the Court sought to avoid when instructing the parties to exhaust collaborative efforts to find a mutually agreeable resolution. ECF 715 at 3. Instead of

[1] Just this week, the latest OpenAI witness to testify to this point, Qiming Yuan, testified that [redacted] (**Exhibit B**, Yuan Rough Tr. at 90:6-10) and [redacted] (*id.* at 72:16-18).

[2] These include LibGen, Internet Archive, Z-Library, Common Crawl, WebText, Project Gutenberg, Anna's Archive, Open Library, Reddit, DuXiu, and E-Libra. ECF 546 at 1 n.1.

submitting a supplemental brief to address burden and proportionality issues, Class Plaintiffs use this latest "renewed" motion (ECF 777) to seek broader and more unwieldy discovery. They now demand inspection of **both** the Bingdex tool **and** the current Bing Index to search for **any books** that may be contained therein. They seek to conduct searchers for books regardless of whether such books are copyrighted, or whether they were used in model training. The demand is thus no longer cabined to the contents of the current Bing Index, nor is it even limited to the eleven originally-identified "notorious piracy websites."[3] ECF 777 at 1. Not only is the substantial burden on Microsoft in seeking even more expansive discovery obviously not alleviated, as articulated in more detail below, it is also a far cry from the "compromise" that the Court directed the parties to pursue. Microsoft has addressed the Court's proportionality concerns by virtue of the tailored discovery proposal communicated by Microsoft, and respectfully requests that this Court again deny Class Plaintiffs' motion to compel production and/or inspection of the Bing Index.

## I. Inspection of the Bing Index is Unduly Burdensome and Disproportionate, Especially When Microsoft is Producing a Much More Tailored Set of Information.

Maintaining the Bing Index in the ordinary course is a massive undertaking at baseline. As exhaustively explained by Microsoft and its 30(b)(6) witness, the live Bing Index is constantly churning as it is dynamically updated. *See* ECF 587 at 3; *see also, e.g.*, ECF 546-2 at 156:25-157:2, 158:23-159:3. While the Court has already found an inspection of much of the content of the live Bing Index to be of "limited relevance," if Class Plaintiffs nevertheless still wish to query it, they can readily do so by using the publicly-available search function on Bing.com, as the Bing Index is a collection of searchable, available web pages crawled by the Bingbot crawler.

Because the Bing Index is "continuously changing every minute" (ECF 546-2 at 20:2-16), it simply cannot be inspected like a traditional training dataset without having to constantly revisit each URL over and over and over again to account for continuous changes. **Exhibit C**, Canel Decl., at ¶ 3. The alternative—creating a static copy of the Bing Index for inspection—would be overwhelmingly burdensome. Due to the magnitude of content on the internet, the compressed size of the current Bing Index approaches [redacted] Canel Decl. at ¶ 3; *see also* ECF 546-2, at 42:23-43:9.

Given its enormous size and complexity, a custom pipeline would have to be built to export such a static copy to a storage location for inspection. This build would require weeks of developer and publisher time to write the code and prepare the destination location for publication. Based on Microsoft's investigation, this substantial effort would take at least weeks, if not months, to construct and would cost several million dollars in construction and storage costs. Canel Decl. at ¶¶ 4-6; ECF 546-2 at 77:6-20; 93:4-9 (testifying that it took several weeks over the holidays for a team of developers to build a pipeline for prior transfer).

Given that "the parties do not appear to dispute that the near entirety of the Bing Index is irrelevant to this litigation," (ECF 712 at 9), Class Plaintiffs' new demand stretches far beyond the bounds of what is necessary to facilitate the discovery as originally requested and with blatant disregard to relevance, or the central question of whether the burden or expense of the proposed

[3] This sweeping request appears to be based on the purported existence of LibGen URLs in the Bing Index. To the extent that LibGen URLs are indeed the scope of what Class Plaintiffs seek to review, then Microsoft has already agreed to provide this information to Plaintiffs in the proposal attached hereto as Exhibit A, and Class Plaintiffs' current motion is moot.

discovery outweighs its likely benefit.[4] This is why Microsoft made the proposal set forth in Exhibit A: producing content from URLs identified through querying Microsoft's Bingdex tool.

## II. Class Plaintiffs Did Not Comply with This Court's Order to Cooperate to Find Ways to Alleviate Burdens on Microsoft in Light of the "Limited Relevance" of Today's Bing Index.

When this Court denied Class Plaintiffs' motion to compel production of the Bing Index without prejudice, this Court directed the parties "to continue meeting and conferring on this issue, particularly with respect to the technological burdens of review and/or production on Microsoft, *as well as ways to alleviate such burdens*." ECF 715 at 3 (emphasis added). The Court specifically encouraged the parties to "*think about whether there is something more limited or something that maybe Microsoft has ideas about ways to alleviate the burden*, whether it's on them or on you. Because my hope was that my decision, at least on relevance and setting out some concerns about relevance and about proportionality, gives you some guidance about where you to go next, okay?" Oct. 29 Hearing Tr. at 33-34 (emphasis added).

Class Plaintiffs assert that "Microsoft has offered nothing" in response. This could not be further from the truth. Exhibit A sets forth a reasonable proposal to collect the contents from each of the domains identified by Class Plaintiffs but excludes Reddit and Internet Archive due to their sheer volume and because they are not "notorious piracy websites" by any reasonable standard.[5] Nor was this "first proposed" the eve before briefing—the parties already conferred regarding this approach on November 7, and reiterated this proposal in the November 7 Joint Status Letter to the Court (ECF 731 at 1). Microsoft reduced this proposal to writing on November 17, with even more specificity in an effort to close the book on this dispute. And contrary to Class Plaintiffs' suggestion, the proposal that Microsoft delivered is not a "partial" resolution to the issue—Microsoft offered to collect and produce content from all but two of the requested domains to reach a compromise, notwithstanding its irrelevance to the core issues in this case.[6]

As we approach the December 4th conference, the parties and the Court now have spent months of substantial time and effort regarding discovery into the current Bing Index, which indisputably was never used to train any model in this case. Because it has never been used to train any model in the case, it should come as no surprise that it is not referenced anywhere within the four corners of the operative complaint in this case. Yet here we are.

It is long past time to bring this ongoing discovery saga into the current Bing Index to a close. Microsoft's reasonable discovery proposal set forth in **Exhibit A** does exactly that. Anything more, including Class Plaintiffs' latest, unqualified request to inspect the current Bing Index,

---

[4] Class Plaintiffs rely on an email thread to suggest the notion that the requested discovery is as simple as clicking around in the BingDex tool. However, the core of their request is not for statistics from the BingDex—it is for the pure content stored in the Bing Index associated with each of these websites. Even still, prior to Class Plaintiffs' "renewed" motion, Microsoft independently offered to provide both these statistics and the contents that Microsoft understood as the basis of the request. *See* Exhibit A at 4-5.

[5] Microsoft reasonably proposed to limit that volume by excluding Reddit and Internet Archive—two sites which collectively represent [redacted] in the Bing Index and which have the least likelihood of containing the content sought by Class Plaintiffs. Canel Decl. at ¶ 10.

[6] Throughout these discussions, Microsoft has made clear that it intended to come to a mutually agreeable resolution. In correspondence regarding the Joint Letter filing, Microsoft made clear that "Although we still have a dispute today, we are continuing to investigate the feasibility and relative burdens of potential alternatives, and intend to work with class plaintiffs in an effort to resolve this dispute." **Exhibit D** at 4.

would be both unduly burdensome and wildly disproportionate to the core issues in dispute in this case, and should be denied.

| | |
|---|---|
| Respectfully submitted, | Respectfully submitted, |
| */s/ Annette L. Hurst* | */s/ Jared B. Briant* |
| Annette L. Hurst | Jared B. Briant |

*Counsel for Defendant Microsoft Corporation*