UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE:	:	
	:	
OPENAI, INC.,	:	25-md-3143 (SHS) (OTW)
COPYRIGHT INFRINGEMENT LITIGATION,	:	
	:	
	:	**OPINION & ORDER RE: PRODUCTION OF THE BING INDEX (ECF NOS. 546, 547, 587, 588)**
This Document Relates To:	:	
23-CV-8292	:	
23-CV-10211	:	

ONA T. WANG, **United States Magistrate Judge**:

Pending now before the Court is Class Plaintiffs' motion to compel Microsoft to produce or make available for inspection the Bing Index. (*See* ECF 546).[1] For the reasons set forth below, Class Plaintiffs' motion is **DENIED without prejudice.**

I. BACKGROUND

The Court assumes familiarity with the facts of this case and summarizes the factual and procedural history relevant to the pending motion.

As part of the normal operations of its Bing Search Engine, Microsoft creates and maintains what it calls the "Bing Index," an internal repository of uniform resource locators ("URLs") and, in most instances, text data associated with those URLs. (*See* Sept. 25, 2025, Transcript, ECF 693 at 69-75) (hereinafter "Sept. 25 Tr."). The Bing Index is not a simple dataset representing a snapshot of the URLs at a point in time—it is a dynamic repository that is "changing every minute." (*Id.* at 70). Namely, the Bing Index, in scouring the internet, adds new URLs, removes old URLs, and updates existing URLs to keep search results accurate. (*Id.*).

[1] Unless otherwise indicated, all references to the docket refer to 25-MD-3143.

At various points in time between 2019 and 2024, Microsoft shared portions or the entirety of the Bing Index with OpenAI. (*Id.* at 76-77). Between August 2019 and January 2020, Microsoft sent an unknown number of "small-scale drops," which Microsoft's counsel estimates to contain somewhere in the range of several thousand URLs. (*See id.* at 76). *See also* Ex. A at 7.[2] Then, in approximately January 2022 to July 2023, Microsoft shared several "larger-scale transfers" of the Bing Index with OpenAI—these transfers appear to have been large portions of, if not the entirety of, the Bing Index as it was being refreshed. (*See* Sept. 25 Tr. at 76); Ex. A at 7. Microsoft collectively refers to the data transfers that took place between August 2019 and July 2023 as the "Bing Index Data Transfers," which were supposedly governed by the "2019 Bing Index Data Scope of Work" Agreement," (the "2019 SOW Agreement") (Sept. 25 Tr. at 76); Ex. A at 7. Microsoft asserts that the 2019 SOW Agreement limits the use of data shared under the agreement to "computational analysis." (Sept. 25 Tr. at 76).[3]

A second period of data transfers with OpenAI took place between October 2023 and July 2024, which the parties refer to as the "GPTBot Web Crawl" period. (*Id.* at 77); Ex. A at 7. Data transfers during this period were governed by a separate data-sharing agreement—the "2023 Web Crawling Assistance Agreement," (the "2023 WCA Agreement"). (Sept. 25 Tr. at 77). Under the 2023 WCA Agreement, Microsoft and OpenAI both crawled the web for URLs, filtered them, and produced a separate dataset called the "GPT Bot Dataset." (*Id.*). The GPT Bot

[2] Counsel for Microsoft presented a slide deck in support of their arguments at the September 25, 2025, discovery status conference. The slide deck is attached to this Opinion & Order as Exhibit A.

[3] During a deposition, Microsoft's deponent testified that "computational analysis" is limited to experimentation and "analysis," and that such terms are not ordinarily interpreted to include training. (*See* ECF 546-2 at 26). Class Plaintiffs dispute this and argue that the 2019 SOW Agreement does not prevent OpenAI from using data acquired under the agreement to train its LLMs. (Sept. 25 Tr. at 84).

Dataset is a new, cooperatively created dataset that is separate and distinct from the Bing Index, which is maintained solely by Microsoft. (*Id.*).[4] The parties have not apparently raised any discovery disputes regarding this dataset yet.

The SDNY Authors filed their first motion related to the Bing Index on March 31, 2025, before the MDL was convened. (*See* 23-CV-8292, ECF 377). In that motion, the SDNY Authors sought an order compelling Microsoft to produce "all data [that] it scraped and provided to OpenAI's [*sic*] for use or potential use in its LLMs." (*Id.* at 3). The Court heard oral argument on the motion at the May 27, 2025, discovery status conference, after the MDL was convened, and granted Class Plaintiffs' motion in part, directing the parties to meet and confer on whether OpenAI or Microsoft had possession, custody, or control of the information sought. (*See* ECF 78). As the dispute remained unresolved after the parties' meet and confer, the Court again heard oral argument on June 25, 2025, denied Class Plaintiffs' motion as premature, and directed the parties to meet and confer again as to whether the dispute could be resolved through a 30(b)(6) deposition. (*See* ECF 270).

Unsurprisingly, the 30(b)(6) deposition only fueled the dispute. Class Plaintiffs renewed their motion on August 9, 2025, asserting that the current version of the Bing Index contains upwards of 230,000 URLs tied to Library Genesis ("LibGen"),[5] and seeking to compel Microsoft

[4] Microsoft asserts that the GPT Bot Dataset was not used to train the GPT models at issue in this case. (*See* 23-CV-8292, ECF 386 at 2). However, as Judge Stein's recent Opinion & Order on Microsoft's motion to strike the Class Plaintiffs' consolidated class action complaint makes clear, the models that the GPT Bot Dataset was potentially used to train are at issue. (*See* ECF 707 at 6); (*see also* 23-CV-8292, ECF 377-6 at 4) (describing using the GPT Bot Dataset to train GPT-4o).

[5] Library Genesis is a "notorious shadow library which has been repeatedly ordered shut down by this court due to piracy." (*See* ECF 368 at 1 n. 1).

to produce the Bing Index or to make it available for inspection. (*See* ECF 546).[6] Microsoft filed its opposition on August 12, 2025. (*See* ECF Nos. 587).[7] The Court again heard oral argument on this dispute at the September 25, 2025, discovery conference. (*See* Sept. 25 Tr. at 66-105).

II. LEGAL STANDARD

Federal Rule of Civil Procedure 26(b)(1) permits discovery of "any nonprivileged matter that is relevant to a party's claim or defense and proportional to the needs of the case." The party moving to compel "bears the burden of demonstrating relevance and proportionality." *See Winfield v. City of New York*, 15-CV-5236 (LTS) (KHP), 2018 WL 840084, at *3 (S.D.N.Y. Feb. 12, 2018); *see also New York Times*, 757 F. Supp. 3d at 597. "When broader discovery is sought, the Court should determine the scope according to the reasonable needs of the action, depending on the circumstances of the case, the nature of the claims and defenses, and the scope of the discovery requested." *Int'l Code Council, Inc. v. Skidmore, Owings & Merrill, LLP*, 24-MC-412 (DEH) (VF), 2025 WL 1936704, at *2 (S.D.N.Y. July 15, 2025). "It is within a Magistrate Judge's discretion to decide whether discovery requests are relevant and proportional," *Radio Music License Committee, Inc. v. Broadcast Music, Inc.*, 347 F.R.D. 269, 273 (S.D.N.Y. 2024), and "Rule 26 gives a district court broad discretion to impose limitations or conditions on discovery[,] which extends to granting or denying motions to compel." *Edmar Fin. Co., LLC v. Currenex, Inc.*, 347 F.R.D. 641, 646 (S.D.N.Y. 2024).[8]

[6] The unredacted, sealed version of Class Plaintiffs' motion is available at ECF 546. The redacted, public version is available at ECF 547.

[7] The unredacted, sealed version of Microsoft's opposition is available at ECF 587. The redacted, public version is available at ECF 588.

[8] "A court can … limit the frequency or extent of discovery if the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less

III. DISCUSSION

As an initial matter, the Court addresses Microsoft's concern that the extant motion is a foray into the legality of Microsoft's Bing Search Engine. It is not. Class Plaintiffs are not alleging that Microsoft's operation of the Bing Index to support its Bing Search Engine is a violation of copyright law, (*see* Sept. 25 Tr. at 82-83), and it is well settled law in this Circuit and others that such activities are indeed protected by fair use. *See Authors Guild v. Google, Inc.*, 804 F.3d 202, 229 (2d Cir. 2015) ("Google's unauthorized digitizing of copyrighted-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses."). *See also Kelly v. Arriba Soft. Corp.*, 336 F.3d 811 (9th Cir. 2003) ("Arriba's reproduction of Kelly's images for use as thumbnails in Arriba's search engine is a fair use under the Copyright Act."). What Class Plaintiffs are challenging, however, is the transfer of a copy (or copies of portions) of the Bing Index data to OpenAI, ostensibly for use in training LLMs or some other purpose, which Class Plaintiffs allege is a separate use of that copyright-protected data that would potentially go beyond the recognized limits of those fair use decisions.[9]

Next, Microsoft suggests, in passing, that past versions of the Bing Index are not relevant to the case because they were never meant to be or actually were used to train OpenAI's LLMs. (*See* ECF 587 at 3). However, Class Plaintiffs have put forth some evidence to dispute this assertion. (*See* ECF 546-7 at 5). Moreover, I have held that whether potentially pirated copies of books were sent to OpenAI is relevant even if such copies were not used for

expensive." *Delta Air Lines, Inc. v. Lightstone Group, LLC*, 21-MC-374 (RA) (OTW), 2021 WL 2117247, at *2 (internal quotation marks omitted).

[9] Microsoft asserts that the Bing Index was "streamed to an Azure storage location," meaning that there was no "copy" in Microsoft's files that could have been retained. (*See* ECF 587 at 2).

training. *In re OpenAI, Inc., Copyright Infringement Litigation*, --- F. Supp. 3d ---, 2025 WL 2691297, at *3 (S.D.N.Y. 2025) ("The relevant 'uses' at issue in this case are: (1) Defendants' alleged copying of NYT and other plaintiffs' copyrighted works to (a) create an internal dataset and (b) train GPT models…."). *See also Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1022 (N.D. Cal. 2025) (analyzing as a separate use those copies of plaintiffs' copyrighted works that Anthropic retained "in its central library for other uses that might arise *even after deciding it would not use them to train any LLM (at all or ever again)*") (emphasis in original). Judge Stein's recent Opinion & Order on OpenAI's motion to strike portions of the consolidated class action complaint confirms that the Plaintiffs' claims in this case are not limited only to training. (*See* ECF 707 at 1) ("OpenAI's request to strike allegations related to the so-called "download claim" is denied because the claim for relief and factual allegations underlying a download theory of infringement were present in the underlying class action complaints before consolidation."). Thus, if Microsoft still possessed such copies of the data that was shared with OpenAI, whether intended to be used for training or some other purpose, those copies would certainly be relevant.

Nevertheless, the data that was shared from the Bing Index between 2019 and 2023 is no longer available,[10] and the only dataset currently implicated by Class Plaintiffs' motion is the current version of the Bing Index.[11] Microsoft argues that the current version of the Bing Index

[10] At the September 25 conference, Class Plaintiffs hinted that Microsoft's and/or OpenAI's deletion of the data transfers from the Bing Index was "prejudicial," (Sept. 25 Tr. at 89), which inches towards spoliation. The Court will not address spoliation here, and Class Plaintiffs indicated at the conference that they were still engaged in discovery on this potential issue. (*Id.* at 105).

[11] It is unclear from Class Plaintiffs' arguments at the September 25 conference whether they seek data transfers in 2023 that would implicate data transferred under the 2023 WCA Agreement. (*See* Sept. 25 Tr. at 67) ("We would love to look at the … data drops that were provided in 2019, 2020, 2022, and 2023."). Because Class Plaintiffs'

is not relevant because it was not used to train OpenAI's LLMs, and even if it were relevant, it is a poor proxy for prior versions because of how frequently the Bing Index is updated (i.e., new URLs are added, current URLs are refreshed). (ECF 587 at 3).

Even if the current version of Bing Index is not used to train OpenAI's LLMs, its relevance is derived from its ability to be used as a proxy for the clearly relevant past versions. Although there are certainly differences between the past versions of the Bing Index and the current version,[12] Microsoft's 30(b)(6) deponent testified that Microsoft could use an internal tool called the "Bingdex" to search the Bing Index for metadata associated with certain URLs currently in the Bing Index, including when a particular URL was crawled and added. (*See* ECF 546-2 at 8). This means that if a URL related to LibGen is currently in the Bing Index, Microsoft can use Bingdex to determine whether that link was present during the relevant data transfer time period. (*See* Sept. 25 Tr. at 94).

Even so, this does not necessarily mean that the URLs present today were accompanied by text data when they were shared with OpenAI. As Microsoft explained at the September 25 conference, several layers of protection exist that prevent the inclusion of URLs and text data related to pirated content. First, Microsoft utilizes robots.txt files: if a site is "robots blocked by the webmaster [or the URL]," the robots block is honored, and the URL is not added to the Bing Index, meaning no associated text data will be present either. (Sept. 25 Tr. at 72). Second, even for URLs without a robots.txt file, Microsoft asserts that the Bing Index cannot process

motions focus exclusively on the Bing Index, the Court assumes that Class Plaintiffs' motion is limited to data shared between August 2019 and July 2023 under the 2019 SOW Agreement. *See* Ex. A at 7.

[12] (*See* ECF 587 at 3) ("There are 230,000 LibGen links in the current Bing Index. My question is, how many LibGen links were in the Bing Index that Microsoft shared with OpenAI in 2023? A: I have no way to know this.") (internal quotation marks omitted).

torrented links, so the text data for any pirated copies of Class Plaintiffs' works associated with these URLs should not be included in the Bing Index. (*Id.*). Class Plaintiffs do not dispute these assertions, and implicit in their argument is the assumption that if they searched a LibGen link in today's Bing Index using the Bingdex and discovered that it was added during the relevant data-sharing period, that necessarily means that text data for those URLs was also shared with OpenAI. But for those reasons outlined above, that may not be true. It is not readily apparent that the date that a URL was added to the Bing Index alone can tell us whether there also was text data associated with that URL at that time, or whether such text data was actually shared with OpenAI.[13] This does not mean that there is no possibility that past versions of the Bing Index contained text files of Class Plaintiffs' copyrighted works and/or that such text data was ever shared with OpenAI; however, it does suggest that the current version of the Bing Index is a flawed proxy for the version(s) of the Bing Index that were shared between August 2019 and July 2023, which weakens its relevance.

The burden of production here is also significant given the limited relevance of the current Bing Index. The Bing Index is vast in size (upwards of one trillion URLs)[14] and constantly "churning," yet Class Plaintiffs request that Microsoft be compelled to use the Bingdex to search each of the potentially 230,000 LibGen-related links currently in the Bing Index to determine whether (1) the current links contain text data for Class Plaintiffs' copyrighted works and (2) the link was crawled during the relevant period. (*See* ECF 546 at 3). Class Plaintiffs have

[13] (*See* Sept. 25 Tr. at 91) (quoting Microsoft's 30(b)(6) deponent's testimony: "we do observe that there were some URLs in the index. What we cannot tell is whether the content of those URLs was present in the index or not because a lot of times we can have URLs but not necessarily the content").

[14] (*See* Sept. 25 Tr. at 69).

not represented that their request is narrower than a full-scale review of the entirety of the 230,000 LibGen-related links and only ask in the alternative that Microsoft produce the entirety of the Bing Index—the parties do not appear to dispute that the near entirety of the Bing Index is irrelevant to this litigation. (*Id.* at 2 n. 2) ("To the extent Microsoft considers it too burdensome to review the Bing Index and only produce data from relevant Shadow Libraries, Plaintiffs are amenable to Microsoft producing the full Bing Index and to running those searches themselves.").[15] Given the limited relevance of today's Bing Index based on currently available information, both may be too burdensome.[16]

At this stage, there is insufficient information to justify either a review of all LibGen-related links currently in the Bing Index by Microsoft or a full-scale production of the entirety of the Bing Index to Class Plaintiffs for their own review of such links. However, Microsoft has also failed to articulate the burden associated with such review or production (other than asserting that one exists), and thus there is insufficient information to show that such discovery is not proportional. The Court will address the proportionality of Class Plaintiffs' request at tomorrow's discovery status conference.

IV. CONCLUSION

For the foregoing reasons, the Court **DENIES** Class Plaintiffs' current motion to compel **without prejudice**.

[15] (*See also* Sept. 25 Tr. at 95) ("It is so easily tested in discovery. Let's investigate. Let's see. Let's allow the inspection to see if the data crawled from 230,000 LibGen sites contain copyrighted books or not.").

[16] Class Plaintiffs do not appear to have not considered, and it is not clear to the Court at this time, the potential technological difficulties that full-scale production of the Bing Index might entail for Microsoft. Would this require the creation of a new Azure environment and an external Bingdex tool, similar to Microsoft's arrangement with OpenAI? Or something else? Without this information, the Court cannot assess the proportionality of Class Plaintiffs' request that Microsoft produce the entirety of the Bing Index.

The Clerk of Court is respectfully directed to close ECF Nos. 546 and 547.

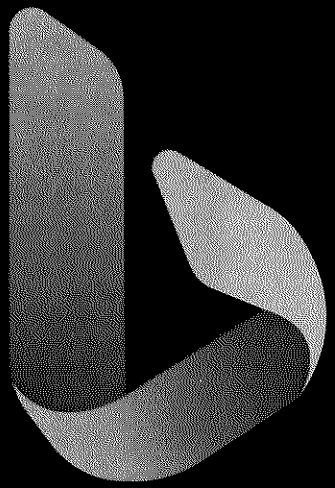
SO ORDERED.

s/ Ona T. Wang

Dated: October 28, 2025
New York, New York

Ona T. Wang
United States Magistrate Judge

EXHIBIT A



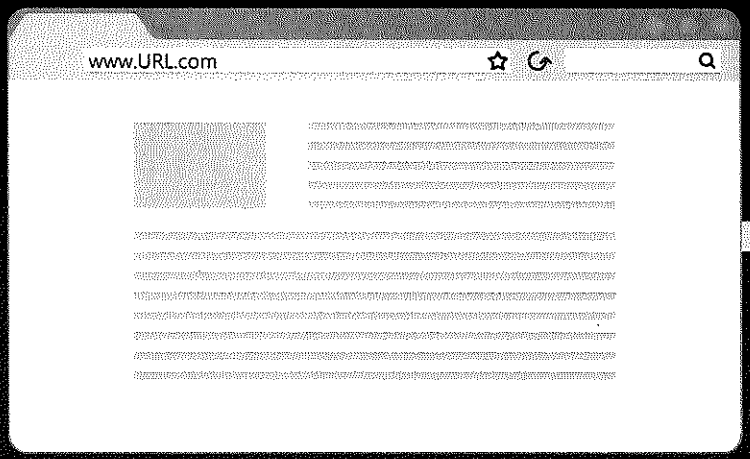
The Bing Index

June 2009

The Bing Index Constantly Refreshes Webpages

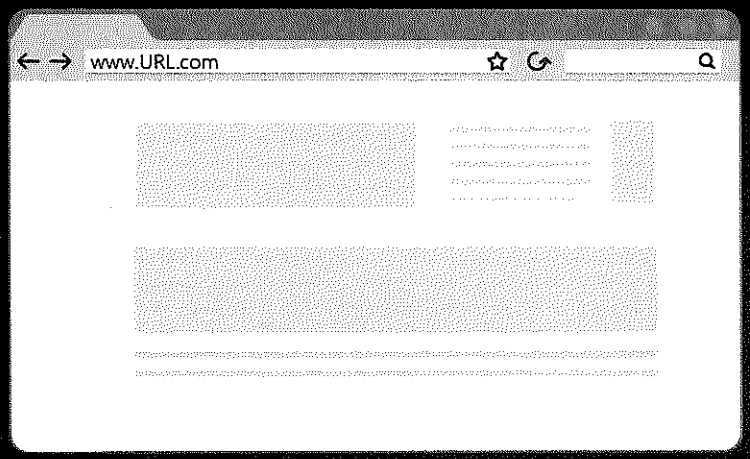
www.URL.com

(January 2022)

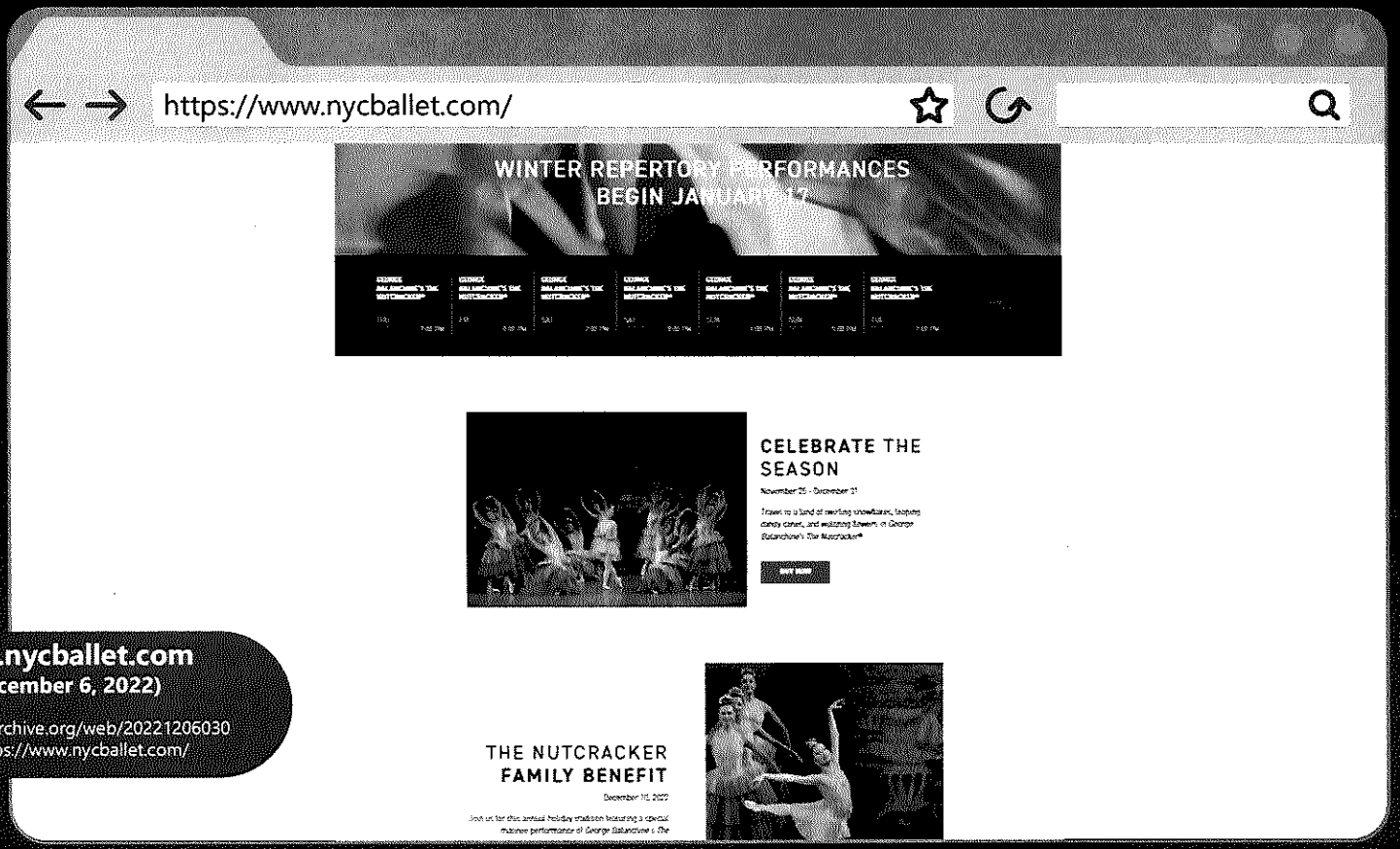


www.URL.com

(September 2025)



An Example



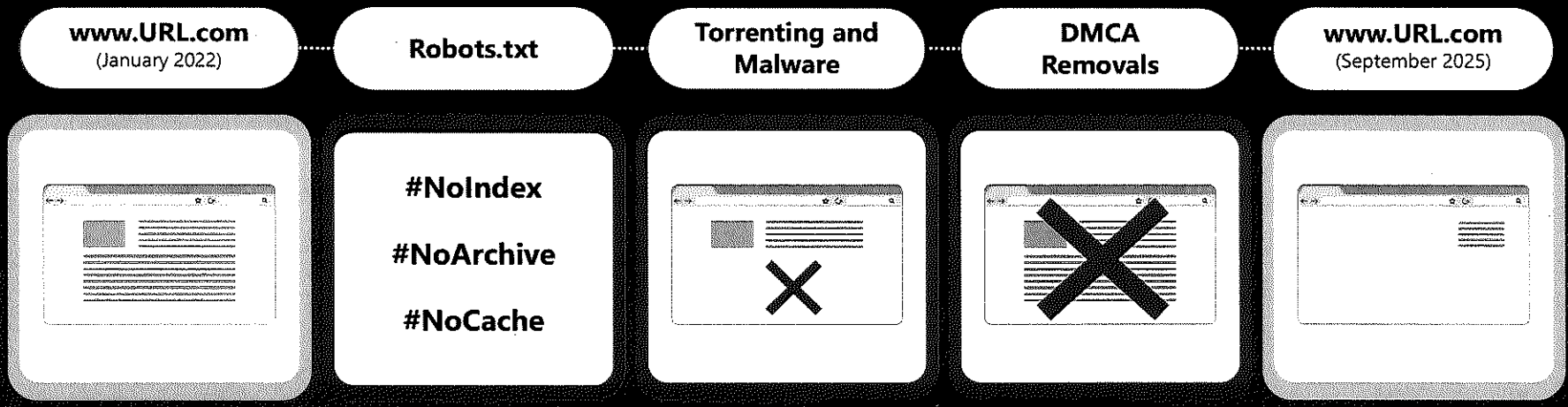
www.nycballet.com
(December 6, 2022)

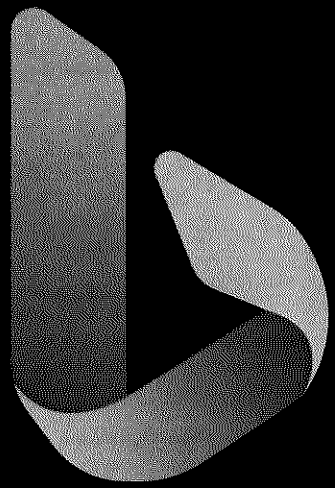
https://web.archive.org/web/20221206030832/https://www.nycballet.com/

An Example

www.nycballet.com
(September 24, 2025)

The Bing Index Dynamically Filters Webpages





The Bing Index

August 2019 – July 2023

Bing Index Data Transfers

2019 Bing Index Data SOW
Experimental Purposes

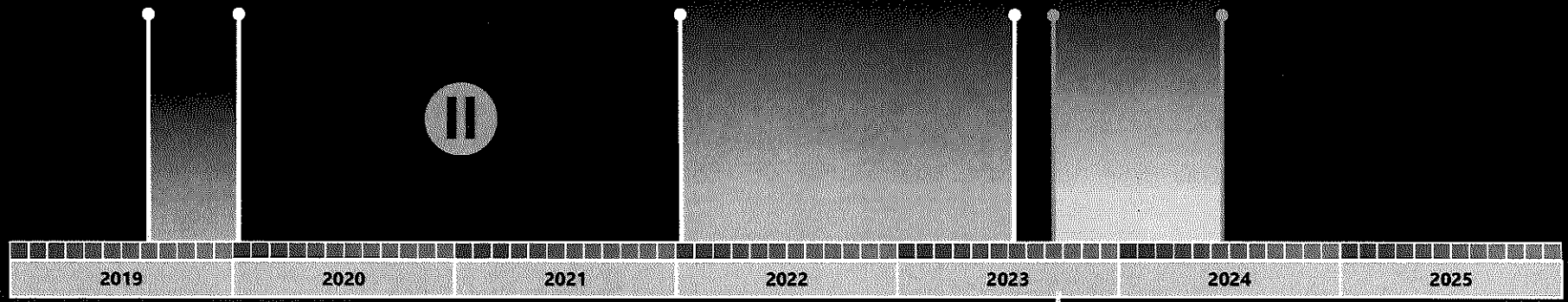
Aug. 2019 – Jan. 2020
Small-Scale Drops

Jan. 2022 – Jul. 2023
Larger-Scale Transfers

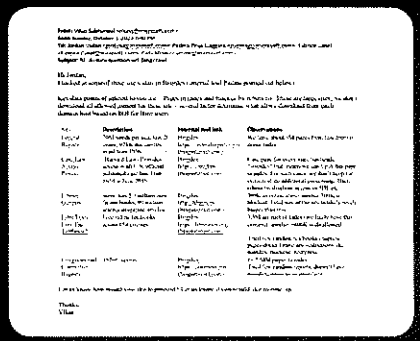
GPTBot Web Crawl

2023 Web Crawling Assistance Agreement
Web Crawling Services for OpenAI

Oct. 2023 – Jul. 2024
Web Crawl Services



Oct. 11, 2023
Exploratory List of AI Training Data Sources



Bing Index Data Transfers

2019 Bing Index Data SOW
Experimental Purposes

GPTBot Web Crawl

2023 Web Crawling Assistance Agreement
Web Crawling Services for OpenAI

Aug. 2019 – Jan. 2020
Small-Scale Drops

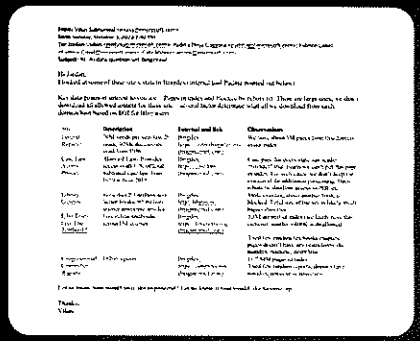
Jan. 2022 – Jul. 2023
Larger-Scale Transfers

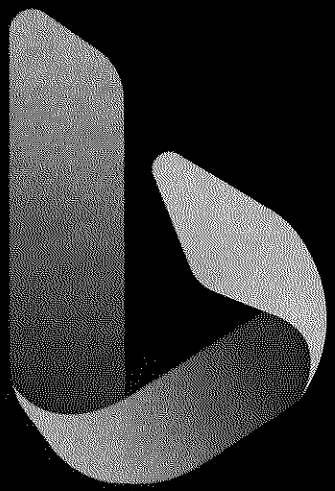
Oct. 2023 – Jul. 2024
Web Crawl Services

September 2025
Current Bing Index

2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025

Oct. 11, 2023
Exploratory List of AI Training Data Sources





The Bing Index

September 2025