[redacted]

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: Case No. 1:23-cv-11195 Case No. 1:24-cv-03285 Case No. 1:24-cv-04872	Case No. 1:25-md-3143-SHS-OTW

DECLARATION OF ELBIO ABIB IN SUPPORT OF MICROSOFT CORPORATION'S MOTION FOR SUMMARY JUDGMENT

I, Elbio Abib, hereby declare as follows:

A. Personal Background

1. I am a Corporate Vice President for Outlook at Microsoft Corporation ("Microsoft") and submit this declaration in support of Microsoft's Motion for Summary Judgment. Unless stated otherwise, all facts stated herein are within my personal knowledge. If called upon, I would be willing to testify as set forth in this declaration.

2. I have been employed at Microsoft since 2005, serving numerous roles focused on the development of Microsoft's search products (including the Bing search engine). I have been leading Microsoft's innovation of AI-powered tools since May 2026. I have a Bachelor's degree in Computer Science from the Federal University of Rio Grande do Sul and a Master's degree in Informatics and Optimization from Pontifícia Universidade Católica do Rio de Janeiro.

3. During my years at Microsoft, I have been centrally involved with the development and improvement of the Bing search engine. I have led major portions of

[redacted]

Microsoft's Bing organizations, overseeing large global teams responsible for search quality, user experience, and machine learning integration. More recently, I was part of the core team that developed Bing Chat (now known as Copilot), which integrated OpenAI's GPT-4 large language model technology with Bing search functionalities. Specifically, my responsibilities included developing the "web grounding" capabilities for Copilot. As further discussed below, those capabilities integrate external information retrieved from reliable and authoritative web search results, enabling Copilot to provide relevant responses with increased accuracy.

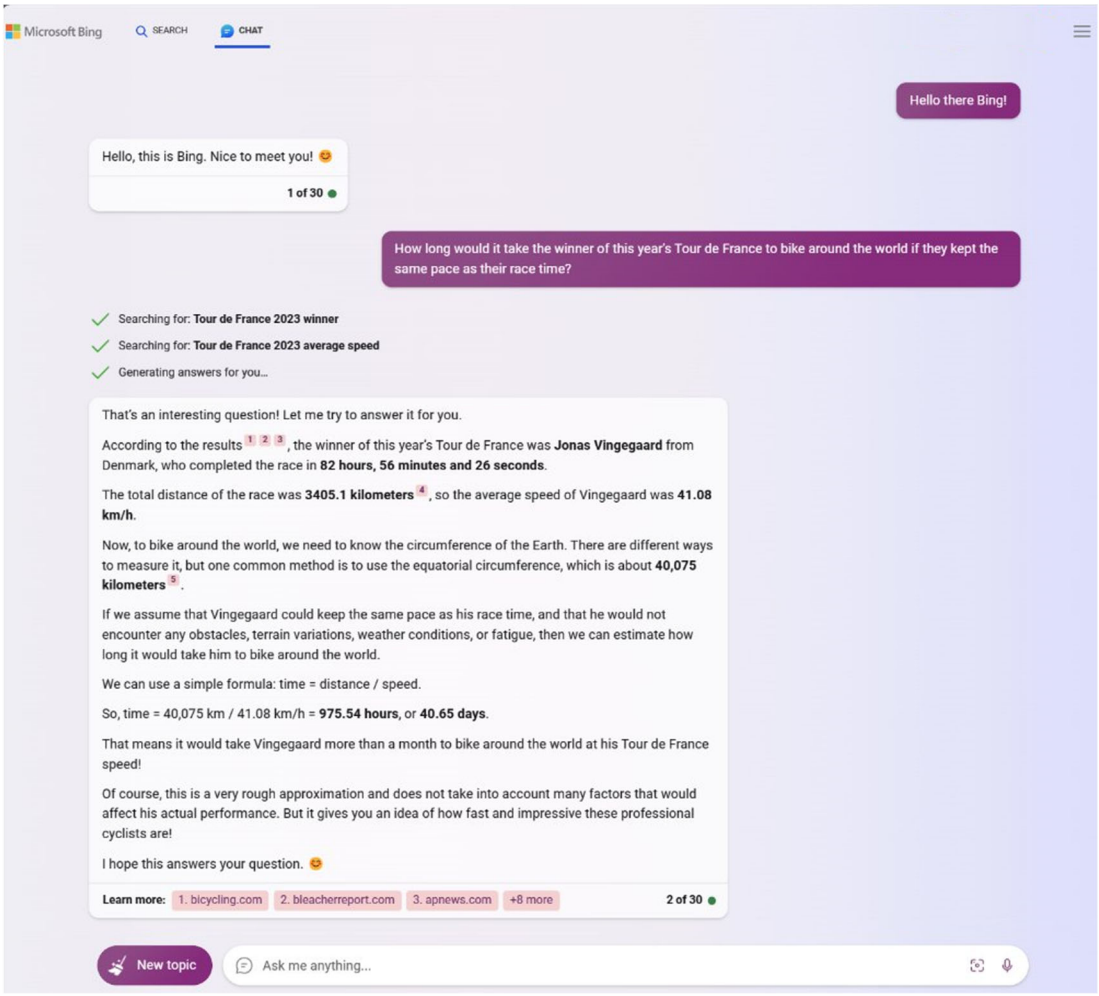
4. In November 2025, I sat for a deposition in this case both as an individual and as Microsoft's designated corporate witness under Rule 30(b)(6). My testimony addressed, among other topics, the technical aspects of Copilot's grounding capabilities.

B. Overview of Microsoft Copilot

5. Copilot is an all-purpose generative AI system designed to assist users with a broad, dynamic spectrum of tasks and across a wide range of topics. By combining advanced large language models with retrieval capabilities, Copilot can help users not only find information but also interpret, synthesize, transform, and act on it.

6. When a user seeks out specific information from Copilot, Copilot is intended to provide tailored responses that synthesize reliable and verifiable sources of information derived from web search results. These responses include linked citations to webpages, enabling users to view and easily access the sources of information that Copilot used to ground its response. Users can click those links to go directly to a source if they want to learn more. Users can also engage in a dialogue with Copilot to find additional information and deeply explore the topic of interest through a conversation that is uniquely customized for that specific user's request. For example:

[redacted]



Copilot's responses in this regard synthesized information across numerous sources that ordinarily could not be returned in one-shot search results using a single query.

7. In addition to learning and knowledge discovery through Copilot responses grounded in web search results, customers use this technology in numerous other ways. For example, the Microsoft 365 Copilot user base commonly leverage this AI assistant for tasks that enhance productivity, such as quickly locating internal company information, problem solving, document analysis, writing assistance, amongst many others. Consumer Copilot users also leverage these productivity enhancing capabilities, as well as numerous other categories of tasks, such as creative content generation, brainstorming, and

[redacted]

transforming content, to name a few. The potential use cases for Copilot are virtually infinite and continue to expand as users learn how to interact with generative AI and optimize its capabilities to suit their unique tasks.

C. Copilot and Retrieval-Augmented Generation

8. Copilot uses a technique commonly referred to as "grounding" or "retrieval-augmented generation" ("RAG"). Grounding refers to the process of providing a large language model ("LLM" or "model") with information from external sources so that the model can generate a response based on that information, rather than relying solely on its predictive language capabilities. One potential external source for such information is the internet. When Copilot grounds responses in web content from Bing search results, this is called "web grounding."

9. RAG techniques provide Copilot with significant advantages over using LLMs alone and greatly improve users' experiences. In particular, LLMs such as those used by Copilot are trained on massive and diverse corpuses of data and learn statistical relationships between words and concepts. Although such models can generate coherent responses, they are incapable on their own of leveraging external sources of information when necessary to generate a response consistent with the user's expectations. Copilot addresses this limitation by combining two capabilities: (1) "retrieval" to search for selected data sources for information relevant to the user's question (*e.g.*, files, emails, chats, meetings, shared workspaces, or web search results); and (2) "augmented generation" to provide the retrieved information along with the user's prompt as context for the LLM to generate an answer. This technique does not change the base model's weights or parameters; instead, it is an additive feature of the Copilot experience that operates at run time.

[redacted]

10. Grounding Copilot responses in web search results (*i.e.*, web grounding) offers numerous benefits for locating information and improving accuracy.

a. First, web grounding allows the LLM to access information that would otherwise be unavailable to it but may be useful in generating the best response. Because information on the web changes continuously (due in part to the constant feed of newly created or freshly updated content), web grounding allows the system to generate answers based on up-to-date and accurate information.

b. Second, Copilot outputs include citations and links to each of the web search results used to inform its answer. This design feature points users to directly relevant web results and enables them to further explore the information they seek.

c. Third, Copilot requests the LLM to tie each informational assertion in its response to specific sources, which reduces the risk of it generating unfounded assertions. In other words, grounding reduces the chances of the model generating a response, in full or in part, that often appears plausible but is unsupported by credible and accurate information (sometimes called "hallucinations").

d. Finally, Copilot responses provide source attribution using citations and links. This feature enables users to verify the accuracy of Copilot's responses themselves, as they are specifically encouraged to do through clear notices that AI may make mistakes. Although Microsoft continuously works to improve Copilot's design to prevent ungrounded answers (an unintended outcome), a perfect success rate is impractical because, at the end of the day,

[redacted]

some error rate, no matter how marginal, is to be expected with any AI technology.

a. Technical Process for Grounding in Web Search Results

11. The foundation of Copilot's generative AI capabilities is its backend technology. Copilot is not itself an LLM. Rather, Copilot integrates LLM capabilities into a system of microservices that manages the core chat capabilities (an orchestration layer known as the "Sydney Orchestrator"). The Sydney Orchestrator includes, for example, microservices for interpreting the user's response, identifying and calling tools and plugins, selecting the appropriate LLM, and integrating Responsible AI technology, amongst numerous other technical components.

12. Response generation is an automated and near-instantaneous algorithmic process that begins when a user submits a prompt to Copilot. The prompt may request that Copilot perform a task, answer a question, or provide information.

13. First, the Sydney Orchestrator analyzes the user's prompt to determine the user's intent—that is, the nature of the requested task. Before the system even begins the process of generating a response, it runs a series of Responsible AI "classifiers" to detect user requests that may result in the system generating a harmful or undesirable response.

14. If the Responsible AI classifiers do not detect a potential risk, the conversation proceeds. The Sydney Orchestrator then selects the LLM with the most appropriate capabilities to field the request.

15. The Sydney Orchestrator sends the LLM the user's prompt along with "prompt instructions" (sometimes referred to as a "metaprompt"), which are instructions provided to the LLM to steer the model's behavior. For example, metaprompts direct the model to request a web search when the user seeks information or when additional

[redacted]

information could improve the response. If the LLM determines that a web search is necessary or would be useful, it initiates the "retrieval" step of the grounding pipeline and sends a request to the Sydney Orchestrator to initiate a Bing web search.

16. The Sydney Orchestrator then composes a search query (or often multiple search queries) derived from user's prompt. Generally, the search queries are short and targeted, rather than submitting the exact prompt the user provided. These search queries are submitted to the Bing search engine endpoint, equivalent to the way a user submits a search query to Bing.com.

17. Initially, Bing retrieves top-ranked search results in the same way that it returns webpages that would appear on the first page of Bing search results. Bing ranks web search results by heavily weighting features such as relevance, quality and credibility, and freshness. Accordingly, Copilot's use of the top-ranked web results ensures that the retrieved web sources are comparatively reliable and authoritative.

18. When a search request comes from Copilot, rather than a traditional search query from Bing.com, the search results are further filtered and reranked then selected before they are sent on to Copilot for the LLM to use for grounding.

a. First, the system filters out any webpages that apply signals indicating that the content should not be used for grounding Microsoft's generative AI answers, *e.g.*, the NOARCHIVE meta tag. The paywall filtration mechanisms for web content sent to Copilot cannot be overridden. This differs from Bing search and related features like Deep Search (a Bing search feature that uses an improved ranking system, also known as "Search Harder"), which may display links and snippets for paywalled content just

[redacted]

as in Bing or in Google. In Copilot, however, any use of such content is the result of a rare, unintended system error.

b. Second, the system applies a logic to the remaining search results that are eligible for use in grounding Copilot answers to determine which webpages contain information most relevant to the user's intent. [redacted]

c. Third, the logic algorithmically re-ranks these results, optimizing for better grounding input to the LLM.

19. Not all content from the selected webpages become potential grounding material for Copilot. [redacted] This limitation always applies to webpages applying the NOCACHE meta tag, which allows Copilot to use only the webpage's URL, title, and snippet for grounding an answer. Even where there are no limitations to Copilot's ability to access the full content of a webpage, the logic may still select only a portion of that webpage based on an algorithmic evaluation of the content most relevant to the user's request.

20. Neither the selection of sources nor the amount of content selected for Copilot grounding reflects an independent determination of a webpage's relative value. Instead, the system selects webpages that are relevant and faithful to the user's request while also honoring the scope of permissions granted by web publishers. Further, the use of full semantic documents for some webpages but not others is due to the LLM's context

[redacted]

window size. In other words, the LLM can only consider up to a certain number of tokens as context (including the user's prompt, Copilot's system instructions, and grounding data) before its generative capabilities begin to deteriorate. This serves as an inherent limitation on the amount of grounding data that may be provided to the LLM, a limitation that is often even tighter when considering that one user query to Copilot may require a combination of information gathered from multiple different web searches. This ranking and selection logic is therefore an algorithmic determination of how to optimize the information for Copilot's grounding use case.

21. The Sydney Orchestrator packages, tokenizes, and transmits the final selection of grounding data to the model as contextual information alongside the user's query and Copilot's prompt instructions. The model draws on the retrieved web content when it generates a response—combining the most relevant points to complete the requested task, ensuring consistency and noting any contradictions, organizing that information in a coherent format, and applying citations to the source that informed each statement.

22. Finally, Copilot displays an output to the user that provides a unique, accurate, and relevant answer to the user's query, expressed in coherent natural language prose. Visible citations and links are appended throughout the response to point users to sources where they can learn more about the information introduced in Copilot's response.

b. Relationship Between Traditional Search and Generative AI

23. Traditional web search engines, such as Google or Bing, and generative AI systems with web grounding capabilities, such as Copilot, share a common foundation—

[redacted]

locating and ranking information on the web to help users obtain information—but differ fundamentally in their purposes, capabilities, and user experiences.

24. The principal purpose of the Bing search engine is to retrieve and rank many existing web pages that may be relevant to the user's query. It returns a list of linked web pages, leaving it to the user to determine which webpages contain useful information and locate useful information within them. By contrast, Copilot is designed for natural language interaction and to produce a coherent, synthesized prose response that directly addresses the user's question. Its output is a readable, self-contained answer that directs users to the sources of information they seek with precision, rather than merely listing places where an answer might be found.

25. Another difference lies in the nature of the way search results are retrieved and presented to the user. Because Bing is optimized for keyword-based queries, users may have to engage in trial and error to formulate an effective search query that will retrieve potentially (but not always) relevant search results. Users may also need to run multiple search queries when searching for information about different topics that may not otherwise surface in a single search result. By contrast, Copilot automates this process using deeply complex algorithms to perform the right searches and find the right information to generate a relevant, synthesized response often in just seconds after the typical user submits a query. Copilot can further engage in multi-turn conversations with the user, maintaining context across a series of exchanges.

26. Accuracy of search results is another key difference. It is acceptable and routine for traditional search engines like Bing to retrieve information that do not track what the user hoped to find. But Copilot requires accurate and authoritative information to responsibly support the claims in its answers.

[redacted]

I declare under penalty of perjury that the foregoing is true and correct. Executed on September 3rd 2026.



Elbio Renato Torres Abib