IN THE 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 KATHERINE COOK IN SUPPORT OF MICROSOFT CORPORATION'S MOTION FOR SUMMARY JUDGMENT

I, Katherine Cook, hereby declare as follows:

1. I am currently a Principal Product Manager within the Turing team at Microsoft Corporation ("Microsoft") with responsibility for M365 Copilot product development. I 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 hold a Bachelor of Science Degree from the University of California, Los Angeles, where I studied cognitive science with a specialization in computing.

3. I have been a Microsoft employee since 2018. From February 2020 through September 2023, I was a Program Manager 2 for Project Turing at Microsoft, at which time I was promoted to Senior Program Manager in Project Turing before being promoted to my current position in September 2025. Through these roles, I am very familiar with

the operation of Bing Chat (later known as Copilot) and Microsoft 365's enterprise version of Copilot, including their web grounding capabilities.

4. I was a product lead for Bing Chat (later known as Copilot). My responsibilities spanned improving system behavior, ongoing product evaluations, integrating safety mitigations and RAI components, and developing the web grounding pipeline (RAG) for Copilot. In addition, I have been working on Microsoft 365's enterprise version of Copilot, including its grounding functionality, since at least mid-2024.

5. Copilot web grounding works by connecting the LLM with Microsoft Bing. When the Copilot system recognizes that responding to a user query may benefit from web grounding, the query proceeds down a "grounding pipeline" as follows:

a) First, the user's query is sent to the LLM to interpret and act upon;

b) Second, Copilot invokes a web search and sends a search-optimized query to the Bing Index, obtaining the top-ranked results;

c) Third, these web results are sent as input tokens to the LLM;

d) Fourth, the web grounding data is placed alongside the user query and prompt instructions as the LLM's temporary "context window";

e) Fifth, the LLM analyzes this "context window" in order to generate a response using its underlying statistical parameters.

6. Copilot's web-grounded responses are influenced by all tokens in the model's context window; the grounded content tokens alter the LLM's probability distribution (its statistical prediction of the next sequence of tokens), which influences the token sequence generated in response to the user query. Copilot is thereby able to synthesize different pieces of web grounding data along with the user prompt and prompt instructions in generating a response, while also drawing upon the background

understanding, insight, and reasoning capacity encoded into its underlying LLM parameters.

7. The Copilot web grounding functionality described above is not designed to display grounding content to a user as part of the LLM's grounded response; that is, the system has no deterministic functionality for passing grounding content through to a user. In other words, grounding content is not copied and pasted into a response shown to a user.

8. Because an LLM's next-token prediction is non-deterministic, an LLM can incidentally predict token strings that match grounding data. It is also possible that an LLM, to satisfy a particular user prompt, would attempt to provide a portion of tokens that match grounding data.

I declare under penalty of perjury that the foregoing is true and correct.

Executed on September 3, 2026.

Signed by:
Katherine Cook
446CA852F19F4A8...

KATHERINE COOK