[redacted]

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: All Cases	Case No. 1:25-md-3143-SHS-OTW

DECLARATION OF JORDAN USDAN IN SUPPORT OF MICROSOFT CORPORATION'S MOTIONS FOR SUMMARY JUDGMENT

I, Jordan Usdan, hereby declare as follows:

A. Personal Background

1. I am the GM, Data Strategy and Ops at Microsoft AI, a division of Microsoft Corporation ("Microsoft"), a position I have held since December 2024. I have been employed by Microsoft since January 2013 in multiple positions. Prior to my current position, I was the Senior Director of Strategy & Innovation at the Office of the Chief Technology Officer ("OCTO"). I submit this declaration in support of Microsoft's Motions 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. In September 2023, I began working on Microsoft's efforts to acquire data for the training of large language models ("LLMs"). In this capacity, I served essentially as air traffic control for data acquisition. My role required me to understand and coordinate the scientific needs of Microsoft's AI teams, established budgets, certain legal

[redacted]

issues, and the business-development work of my colleagues, including Geraldine Russell (current Head of AI Data Partnerships at Microsoft) and her team.

3. Microsoft designated me to be a corporate representative pursuant to Federal Rule of Civil Procedure 30(b)(6) to testify regarding Microsoft's strategy related to the acquisition, purchase, or licensing of data for LLM training and/or grounding and display, among several other topics. Through the positions I have held at Microsoft and over the last few years, I was directly involved in negotiations with book publishers, specifically [redacted] and [redacted], and had visibility into negotiations with other content providers, including other book publishers and news outlets.

B. Microsoft's AI Data Acquisition Strategy and Policies

4. In 2023, OCTO tasked me with coordinating Microsoft's efforts to acquire data for AI-related activities. This involved (1) review of existing data to which Microsoft already had access to determine its suitability for AI-related uses; and (2) exploring and negotiating agreements with third parties to acquire access to data for AI-related uses. Prior to this work, Microsoft did not have paid data access agreements specific to "learning use" (i.e., LLM training). None of the data that Microsoft has acquired through this third-party data acquisition effort was used to train or fine-tune any of the models at issue in this case.

5. Microsoft runs on trust. Microsoft has internalized that key tenet throughout the AI data acquisition process. Microsoft also sees importance in the sustainability of content creation. To that end, for example, Microsoft's Tech for Society initiative focuses on applying technology, including generative AI, and proven processes to secure elections, support journalists, and promote responsible business and

[redacted]

community safety.[1] Microsoft has also approached data acquisition for AI-related uses through this lens of corporate responsibility.

6. [redacted]

1. Microsoft's Third-Party Data Acquisition Efforts

7. [redacted]

[1] https://www.microsoft.com/en-us/corporate-responsibility/trusted-technology-group/tech-for-society.

[redacted]

8. Microsoft produced [redacted] data access agreements with the following entities: [redacted] These entities are all book publishers or academic content providers except for [redacted]. The only book publisher out of the above who publishes trade books is [redacted]. Notably, none of the data to which Microsoft gained access via these agreements was used to train any of the Open-AI-trained models at issue in this case.

9. [redacted]

[redacted]

10. In reality, however, it would be impossible and economically prohibitive for Microsoft to acquire for a fee all of the data needed to train an AI model from a zero token base. The volume and variety of data typically used to train a modern LLM is not available from any one source. It can be obtained only by acquiring it from a wide variety of sources where information can be found on the Internet. There are trillions of tokens available on the web for AI training purposes available to BingBot, which follows opt-out standards and signals for AI training. To make a difference in a training model on top of what is publicly available on the web, Microsoft would need to acquire the equivalent of hundreds of millions of books, more than have ever been written. It would be difficult to detect any difference from even a hundred thousand books added to a training set since such a negligible amount of data represents only a drop in the ocean of the trillions of tokens needed to train an AI model.

11. There is also no accepted or reliable mechanism for valuing how one data input affects AI training. As noted, model development requires an immense number of tokens, and Microsoft does not have any scientific way to determine how one piece of data affects a model (if at all). My team went in circles trying to arrive at a way of valuing data

[redacted]

for AI training, including consulting scientists like Mr. Bubeck, but it is extremely hard to determine the value of a given piece of data is to a model, and then the value, if any, of any increase of performance of that model based on that data input.

a. [redacted]

12. [redacted]

13. [redacted]

14. [redacted]

15. [redacted]

[redacted]

16. [redacted]

b. [redacted]

17. [redacted]

18. [redacted]

[redacted]

19. [redacted]

20. [redacted]

21. [redacted]

22. [redacted]

[redacted]

23. [redacted]

24. [redacted]

[redacted]

25. [redacted]

[redacted]

c. [redacted]

26. Microsoft [redacted]

[redacted]

27. [redacted]

[redacted]

28. [redacted]

[redacted]

[redacted]

29. [redacted]

30. [redacted]

d. [redacted]

31. [redacted]

32. [redacted]

33. [redacted]

[redacted]

34. [redacted] Copilot Daily was a failed experiment and is not in the Microsoft product anymore.

35. Finally, a small experimental program called the Publisher Content Marketplace ("PCM") involves the use of selective news content. The goal of the PCM experiment is [redacted]. PCM is a pilot of a [redacted] PCM is not about accessing data for LLM training. Nor does it relate to books content.

2. Microsoft's Consideration of Existing Data for AI-Related Activities

36. In addition to third-party data acquisition efforts, Microsoft also explored use of data to which it already had access for AI training. Microsoft started by examining the potential efficacy of using data from the Bing Index for AI-related activities. In doing so, Microsoft has implemented two principal rules:

a. First, Microsoft uses content that is publicly available through the Bing Index, subject to web controls that give publishers control over whether and how their content is used by Microsoft. Second, Microsoft does not knowingly train on pirated content.

[redacted]

b. Microsoft's policy since it began training its own models has been to refrain from training on content that Microsoft knows or believes to be pirated, even if that content appears in the Bing Index. Nor does Microsoft seek out pirated content from sources outside the Bing Index.

C. Data Dignity and a Fair and Sustainable AI Data Strategy

37. Microsoft has engaged in good-faith, exploratory efforts on the topic of creator compensation and attribution in the context of AI training data.

38. One such effort involves consideration of the "Data Dignity" effort. Data Dignity predates ChatGPT and arose from the broader concept of compensating many kinds of contributors of data to digital systems prior to generative AI. Data Dignity is not a Microsoft-wide initiative; it is a project conceived by a few individuals employed by Microsoft Research, such as Jaron Lanier, who was hired to bring diverse perspectives to Microsoft, and Glen Weyl. The Data Dignity effort was an incubation to lay out what was referred to as a manifesto and reflected the outside-the-box and idealistic thinking of these individuals. The Data Dignity efforts did not include an executed licensing program for copyrighted material, a mechanism for clearing trade-book rights, or evidence that AI developers could assemble rights for training data at the scale required for general purpose LLMs.

39. Like Jaron Lanier and Glen Weyl, Brent Hecht holds divergent, academic, and forward-looking views about how data ecosystems for AI should operate and is employed at Microsoft to bring these asymmetric, futuristic, and academic points of view. Brent is not someone who sets core priorities for Microsoft, nor is he someone who speaks for Microsoft specifically as to his theoretical views on AI's potential effect on content creators. Brent currently reports to [redacted]

[redacted]

[redacted] Brent is not an economist, but is a research academic with a background in studying data ecosystems. Brent had an academic appointment at Northwestern University concurrent at times with his employment at Microsoft, and he would share the research his Ph.D. students were doing on sustainable AI data strategy with the team at Microsoft. I view Brent's message as urging Microsoft to consider multiple possible data futures and to have a robust, thoughtful data strategy—not as evidence that a mature licensing market already existed. Microsoft Leadership, including OCTO, does not share all of Brent's views, as he is employed at Microsoft to present divergent and asymmetric perspectives.

40. Brent Hecht was not involved in negotiating any of Microsoft's data access agreements or display agreements. Brent never attended any meetings with any third party data partners and was not part of the decision-making process. To my knowledge, Brent does not have access to Microsoft's data access agreements nor has he ever seen them. To any extent that Brent has represented in his testimony or otherwise that he had any involvement in these agreements, any such involvement was peripheral and limited to a few discrete discussions on sociotechnical, big picture items, such as his understanding of the academic publishing business from his experience in academia.

41. The only Microsoft effort that was attempted under the Data Dignity ideology was the Trove app experiment, an internal incubation in which users were asked to complete data collection 'missions,' such as uploading images, in exchange for small payments. The Trove experiment predated the current LLM-training data acquisition work and ultimately failed because it did not develop a sufficiently large two-sided market

[redacted]

of data buyers and data sellers. I am not aware of any other Microsoft experiment implementing Data Dignity beyond Trove.

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

Executed on September 1, 2026.

DocuSigned by:

64E874D220BD421...

JORDAN USDAN