UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

In Re: OpenAI, Inc. Copyright Infringement Litigation This Document Relates To: No. 1:23-cv-08292 No. 1:23-cv-10211	Case No. 1:25-md-3143 (SHS) (OTW) **REDACTED - PUBLIC VERSION**

DEFENDANT MICROSOFT CORPORATION'S RULE 56.1 STATEMENT OF UNDISPUTED MATERIAL FACTS IN SUPPORT OF MOTION FOR SUMMARY JUDGMENT IN BOOKS PLAINTIFFS' CONSOLIDATED CASES

Pursuant to Local Civil Rule 56.1, Defendant Microsoft hereby submits this statement of material facts as to which there are no genuine issues to be tried.

I. COUNT 1: DIRECT COPYRIGHT INFRINGEMENT (17. U.S.C. § 501)

A. Fair Use (17 U.S.C. § 107)

1. *Purpose and Character of the Use (Factor 1)*

1. A large language model ("LLM") is a machine learning deep neural network trained on vast quantities of text with the objective of predicting the next token in a sequence in order to generate natural language.

Lafferty Decl. ¶¶ 16, 33, 77-78; *see also* Hurst Decl. **Ex. 75** (Melton Dep. Tr.) 29:6-30:19, 9:20-10:16, 11:16-17:7; Ex. 112 (Song Dep. Tr.) 32:13-33:25, 9:8-10:21; **Ex. 54** (Huynh Dep. Tr.) 95:23-96:4, 7:11-8:18, 15:3-7, 15:25-16:25; **Ex. 81** (Nadella Dep. Tr.) 274:24-276:4, 277:3-23, 17:19-18:7, 25:16-22, 29:10-30:3; **Ex. 115** (Suleyman Dep. Tr.) 35:9-36:8; **Ex. 99** (Ribas Dep. Tr.) 22:2-25, 10:23-11:6, 11:23-

12:3, 14:2-23, 17:9-18:10, 18:3-19:19.

2. Machine learning requires large amounts of data, and the more complex and varied the domain, the greater the volume and diversity of data needed for a model to learn patterns that generalize reliably to new inputs. Natural language is among the most complex of such domains.

   Lafferty Decl. ¶¶ 16, 41, 45, 52, 66, 70, 77, 81, 90, 97, 99-100.

3. LLMs are a general-purpose technology capable of a wide variety of natural language processing tasks, such as summarization, translation, reasoning, content generation, and question answering.

   Lafferty Decl. ¶¶ 16, 33, 36, 45, 77-78; *see also* Hurst Decl. **Ex. 112** (Song Dep. Tr.) 32:13-33:25; **Ex. 54** (Huynh Dep. Tr.) 95:23-96:4, 7:11-8:18, 15:3-7, 15:25-16:25, 17 :14-22; **Ex. 81** (Nadella Dep. Tr.) 274:24-276:4, 277:3-23; **Ex. 115** (Suleyman Dep. Tr.) 35:9-36:8; **Ex. 99** (Ribas Dep. Tr.) 18:3-19:19, 22:2-25; **Ex. 119** (Templeton 30(b)(6) Dep. Tr. (Vol. 2, dated Jan. 21, 2026)) 407:24-408:9; **Ex. 118** (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 13:14-16, 14:14-20, 15:2-12, 16:18-19:10, 22:1-14, 22:25-23:15, 24:2-25:25, 26:14-27:16, 221:6-222:20.

4. The process of training an LLM involves:

   a) Acquiring a large volume of text to assemble a training corpus—typically by crawling and scraping publicly accessible data on the internet in a manner that has existed before LLMs were invented;

      Lafferty Decl. ¶¶ 69, 79, 89-91, 100; Tucker Decl. ¶¶ 117-122; *see also* Hurst Decl. **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 32:5-33:14, 48:20-24, 66:23-67:10, 95:10-23; **Ex. 130** (Yuan Dep. Tr.)

12:11-14:13, 28:6-12, 8:6-25, 10:9-23.

b) Processing the textual data into a numerical format that can be used by a neural network—*i.e.*, tokenization, whereby readable text is broken into unique "tokens" comprised of short words, segments of longer words, or groupings of characters, and then integers are assigned to each token;

Lafferty Decl. ¶¶ 105-114; *see also* Hurst Decl. **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 19:12-20:18, 176:22-177:23; **Ex. 117** (Sutskever Dep. Tr.) 141:8-142:2, 19:15-19, 21:12-22:7, 26:17-29:8; **Ex. 117.A** (Sutskever Dep. Ex. 12 - Language Models Are Few-Shot Learners GPT-3 Research Paper) 8-11.

c) Embedding the tokens in high-dimensional, mathematical space, where tokens are represented by numerical vectors that point toward and away from each other based on relationships among the tokens gleaned from numerous contexts across the entirety of the training corpus;

Lafferty Decl. ¶¶ 33, 115-123; *see also* Hurst Decl. **Ex. 117** (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 53:3-17, 67:9-15, 141:8-142:2; **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 50:10-19.

d) Processing those vectors through transformer layers repeatedly makes small adjustments to the model "parameters," a set of numerical model weights that the model uses to predict the next token in a sequence of text.

Lafferty Decl. ¶¶ 137-147; *see also* Hurst Decl. **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 18:16-19:4, 176:24-177:23; **Ex. 101** (Rossett Dep. Tr.) 67:25-68:9, 8:1-9, 8:24-9:9, 17:6-18:4, 18:22-23:2; **Ex. 117** (Sutskever Dep. Tr.) 141:8-142:2; **Ex. 117.A** (Sutskever Dep. Ex. 12

- Language Models Are Few-Shot Learners GPT-3 Research Paper) 2-8.

e) Periodically during training, the model will be presented with data it has not seen during training in order to measure the model's generalization ability. Training continues as long as the model's performance on the unseen data continues to improve. Once the model begins to perform well on training data but begins to lose accuracy on testing data, the training stops.

Lafferty Decl. ¶ 71.

5. Training an LLM requires an enormous volume of text. Publicly disclosed training-set sizes for frontier models rose from approximately 2 trillion tokens in 2023 to approximately 18 trillion tokens by late 2024.

Lafferty Decl. ¶¶ 91-92; Hurst Decl. **Ex. 215** (OPCO_MDL_004424223) at -226; **Ex. 197** (MSFT_AI_MDL_000221039) at -042.

6. The GPT models at issue were pre-trained on corpora of this scale: GPT-3 on approximately 300 billion tokens; GPT-4 and GPT-3.5 Turbo on approximately [redacted] tokens (epoch weighted); and GPT-4o on approximately [redacted] tokens (epoch weighted).

Lafferty Decl. ¶¶ 91-92, 94-97; Hurst Decl. **Ex. 196** (MSFT_AI_MDL_000205724); **Ex. 215** (OPCO_MDL_004424223) at -226; **Ex. 197** (MSFT_AI_MDL_000221039) at -042; **Ex. 96** (Radford Dep. Tr.) 20:25-23:3; **Ex. 227** (Paino Decl.).

7. Training an LLM also requires a wide diversity of text. Model developers draw training data from a broad range of sources, including web data, software code, books, encyclopedic and scientific sources, forum posts, and transcripts.

Lafferty Decl. ¶¶ 34, 89-91, 93-97; Hurst Decl. **Ex. 117** (Sutskever Dep. Tr.)

26:16-21, 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; **Ex. 196** (MSFT_AI_MDL_000205724); **Ex. 215** (OPCO_MDL_004424223) at -226; **Ex. 197** (MSFT_AI_MDL_000221039) at -042; **Ex. 198** (MSFT_AI_MDL_000221643); **Ex. 213** (OPCO_MDL_001938121).

8. The internet is the only practicable source of text at a scale sufficiently voluminous and varied to train a general-purpose LLM.

   Lafferty Decl. ¶¶ 45, 90-91; Usdan Decl. ¶ 10; *see also* Hurst Decl. **Ex. 99** (Ribas Dep. Tr.) 80:3-81:23; **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 48:20-24, 67:2-6; **Ex. 117** (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; **Ex. 130** (Yuan Dep. Tr.) 28:6-25, 12:11-14:13.

9. LLM training requires an enormous volume and a wide diversity of textual data because the fundamental objective of training is generalization—producing a model capable of responding to new, previously unseen inputs across an effectively unlimited range of contexts and tasks.

   Lafferty Decl. ¶¶ 16, 17, 34, 90-91, 98, 154-156.

10. The capacity to generalize is an emergent property of learning across the complete distribution of the training data. No individual work is responsible for conferring that capability.

    Lafferty Decl. ¶¶ 16, 33-34, 45, 66, 80-81, 99, 145-146; *see also* Hurst Decl. **Ex. 81** (Nadella Dep. Tr.) 90:9-23, 274:12-275:12.

11. Because the statistical patterns a model learns are distributed redundantly across millions of sources spanning trillions of tokens, the removal of any single work or collection of works from the training corpus would not materially alter the statistical

structure the model learns.

Lafferty Decl. ¶ 33-34, 99, 322.

12. A properly trained LLM is a set of billions or even trillions of numerical parameters (mathematical weights) that encode generalized statistical relationships of language such that the model can simulate conversation and reasoning.

Lafferty Decl. ¶¶ 16, 33-35, 45, 52, 71-72, 74-76, 90-91, 94-100, 139-40, 154-55, 157-58, 164; *see also* Hurst Decl. **Ex. 81** (Nadella Dep. Tr.) 90:9-23, 275:4-276:13; **Ex. 103** (Russinovich Dep. Tr.) 106:4-107:15, 9:3-14, 9:18-23, 12:24-13:15, 14:18-15:17, 16:5-10; **Ex. 16** (Brockman Dep. Tr.) 319:20-320:12, 23:19-23, 24:3-8, 26:22-27:9, 37:9-38:8; **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 50:10-51:23, 181:25-182:4; **Ex. 117** (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; **Ex. 96** (Radford Dep. Tr.) 20:25-23:3.

13. When generalization is strong, an LLM generates coherent and contextually appropriate text about unfamiliar subjects because its parameters represent the deeper structure of language rather than the surface content of the training set; the resulting output is novel text distinct from the source material used for training.

Lafferty Decl. ¶¶ 76-77, 81, 141, 145, 155-56; *see also* Hurst Decl. **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 48:20-50:19, 104:4-17, 180:21-182:4, 184:18-185:18; **Ex. 16** (Brockman Dep. Tr.) 265:19-266:5.

14. Models can at times generate text that was part of their training corpus, a known phenomenon called regurgitation. However, instances of regurgitation in a properly trained model are extremely rare. A model that has a tendency to regurgitate its training data reflects an unintended side-effect of the training process rather than the

intended operation of the model.

Lafferty Decl. ¶ 35; *see also* Usdan Decl. ¶¶ 19-20.

15. Books Plaintiffs' expert Dr. Shan analyzed logs produced by Microsoft of 8.2 million real-world conversations with Microsoft's Copilot in order to assess whether any of those Copilot outputs contained what he qualified as a "regurgitation" of Books Plaintiffs' Asserted Works. Shan's definition of a "regurgitation" was a contiguous near-verbatim block of at least 30 consecutive words (with certain other exclusions).
    a) For 202 of the 212 Asserted Works, Dr. Shan identified no "regurgitations" in the 8.2 million Copilot logs;
    b) Of the remaining 10 works, out of the 8.2 million conversations, Dr. Shan identified a total of 24 conversations meeting his definition of containing one or more "regurgitations," a rate of .00029%;
    c) To arrive at those 24 conversations, Dr. Shan's search code combined the "Text" field, which contains Copilot responses displayed to users, and the "HiddenText" field, which contains web-grounding material not displayed to users; the search did not distinguish between the two;
    d) Shan found no matches of entire books or book chapters when performing his analysis.

    Lafferty Decl. ¶¶ 366-68, 378; Hurst Decl. **Ex. 241** (Shan Opening Report) ¶¶ 40, 156 (24 responses; 10 works; 30-or-more near-verbatim words); Hurst Decl. **Ex. 243** (Shan Reply Report) ¶ 59.

16. At OpenAI's request, Microsoft provided OpenAI with portions of the Bing Index, the web index that powers Microsoft's Bing search engine.

Hurst Decl. **Ex. 118** (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026))

304:9-16; **Ex. 119** (Templeton 30(b)(6) Dep. Tr. (Vol. 2, dated Jan. 21, 2026)) 390:22-391:1; **Ex. 38** (Gaggara 30(b)(6) Dep. (dated Aug. 20, 2025)) 44:13-15; **Ex. 37** (Gaffney 30(b)(6) Dep. Tr.) 131:20-133:2, 9:2-5, 9:15-10:5, 10:19-25; **Ex. 37.A** (Gaffney Dep. Ex. 16) 2-3.

17. The provision of portions of the Bing Index to OpenAI was a step in the process for the purpose of training LLMs.

Hurst Decl. **Ex. 107** (Shan Dep. Tr.) 318:11-319:9; **Ex. 37** (Gaffney 30(b)(6) Dep. Tr.) 131:20-133:2; **Ex. 37.A** (Gaffney Ex. 16); **Ex. 16** (Brockman Dep. Tr.) 323:11-16.

18. Books Plaintiffs' Asserted Works were created to be read, and are sold and licensed to readers for entertainment, information, and edification.

Tucker Decl. ¶¶ 11(b), 48, 52-54; Hurst Decl. **Ex. 35** (Franzen Dep. Tr.) 21:18-22:18; **Ex. 9** (Baldacci Dep. Tr.) 79:7-14; **Ex. 128** (Wolicki Dep. Tr.) 21:1-6, 162:16-163:8.

2. *Nature of the Work (Factor 2) and Amount and Substantiality Used (Factor 3)*

19. The LLM training process deconstructs works into tokens, then analyzes the patterns among various pieces of tokenized content in order to translate those patterns into model weights–all to gain access to underlying facts, ideas, linguistic patterns, and statistical relationships.

Lafferty Decl. ¶¶ 77, 80-81, 105-114, 137-39, 141, 144-47, 283; *see also* Hurst Decl. **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 18:12-19:4, 19:12-20:18, 48:25-49:24, 50:10-19, 104:15-17, 176:22-177:23; **Ex. 101**

(Rossett Dep. Tr.) 18:22-23:2, 67:25-68:9.

20. The patterns extracted through training—including grammar, syntax, semantic relationships, facts, ideas, and concepts—are properties of the aggregate training distribution rather than of any individual work.

Lafferty Decl. ¶¶ 16, 33-34, 45, 66, 80-81, 99, 145-146. 283.

3. *Impact Upon the Market (Factor 4)*

21. Books Plaintiffs are authors whose traditional markets are publication by commercial publishers of books to be read in various formats such as ink on paper, e-books, and audiobooks, as well as the licensing of those commercially published books for translation into other languages and production in other entertainment mediums.

Tucker Decl. ¶¶ 12(a), 25, 48, 64.

22. Books Plaintiffs could not identify any lost sales or licensing opportunities for their Asserted Works as a result of the use of their Asserted Works to develop any of the LLMs at issue.

Tucker Decl. ¶¶ 11(a), 20-46; *see also* Hurst Decl. **Ex. 35** (Franzen Dep. Tr.) 46:5-8, 90:13-24, 91:1-22; **Ex. 87** (Picoult Dep. Tr.) 93:5-16, 99:3-13; **Ex. 39** (Golden Dep. Tr. (dated Mar. 17, 2025)) 210:16-20, 300:7-14, 301:20-302:2; **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 55:23-57:9; **Ex. 128** (Wolicki Dep. Tr.) 148:17-151:20, 188:11-190:1; **Ex. 108** (Shapiro Dep. Tr.) 91:15-92:4, 136:22-24, 223:6-17, 236:19-239:3; **Ex. 9** (Baldacci Dep. Tr.) 224:20-225:10, 233:3-13, 238:1-15; **Ex. 43** (Greer Dep. Tr.) 107:16-108:19, 110:2-19, 113:15-18; **Ex. 73** (Martin Dep. Tr.) 216:22-218:6, 219:12-21; **Ex. 55** (Hwang Dep. Tr.) 149:8-150:24, 151:16-152:25, 153:21-154:12; **Ex. 106** (Schiff Dep. Tr.) 46:23-47:19; **Ex. 44** (Grisham

Dep. Tr.) 176:22-178:11.

23. [redacted]

Tucker Decl. ¶¶ 11(a) 24-37.

24. [redacted]

Tucker Decl. ¶¶ 11(a), 38-40.

25. The only existing and specific economic harm Books Plaintiffs have identified is the loss of payment for the single copy of each work used in LLM training.

Hurst Decl. **Ex. 44** (Grisham Dep. Tr.) 89:8-21; **Ex. 106** (Schiff Dep. Tr.) 44:18-25, 45:11-46:6; **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 55:23-56:9, 58:6-13, 64:4-7; **Ex. 87** (Picoult Dep. Tr.) 99:14-20; **Ex. 128** (Wolicki Dep. Tr.) 149:22-150:5, 174:8-18; **Ex. 9** (Baldacci Dep. Tr.) 223:3-14.

26. Even assuming "market dilution" is relevant to the fourth factor:

a) Books Plaintiffs have identified no instance in which an AI-generated book displaced a sale of any of Books Plaintiffs' commercially published Asserted Works;

Hurst Decl. **Ex. 35** (Franzen Dep. Tr.) 90:13-24; **Ex. 108** (Shapiro Dep. Tr.) 90:15-92:4; **Ex. 43** (Greer Dep. Tr.) 107:16-108:19; **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 47:21-48:6; **Ex. 55** (Hwang Dep. Tr.) 150:10–24; **Ex. 4** (Amer Dep. Tr.) 250:25–251:8; **Ex. 9** (Baldacci Dep. Tr.) 231:5–12; **Ex. 44** (Grisham Dep. Tr.) 187:8–13; **Ex. 87** (Picoult Dep. Tr.) 94:16–23; **Ex. 14** (Branch Dep. Tr.) 73:11–74:1; **Ex. 73** (Martin Dep. Tr.)

219:13–21.

b) Few if any people are using Copilot for the purpose of generating full books;

Amir Decl. ¶¶ 9(b), 47-51; Bail Decl. ¶ 113; Tucker Decl. ¶ 12(d), 83-88; *see also* Hurst Decl. **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 62:17-63:2; **Ex. 60** (Kazi Dep. Tr.) 143:19-144:1; **Ex. 2** (Albu Dep. Tr.) 206:23-25; **Ex. 43** (Greer Dep. Tr.) 145:12-15; **Ex. 97** (Rasenberger Dep. Tr. (dated Feb. 12, 2026)) 282:18–23; **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 62:17-63: 10; **Ex. 4** (Amer Dep. Tr.) 252:17-25.

c) AI-generated books are viewed with disfavor by consumers;

1. Tucker Decl. ¶¶ 12(b), 70-76, 87; *see also* Hurst Decl. **Ex. 4** (Amer Dep. Tr.) 250:25-251:8; **Ex. 87** (Picoult Dep. Tr.) 98:24-99:13; **Ex. 35** (Franzen Dep. Tr.) 90:21-24; **Ex. 55** (Hwang Dep. Tr.) 152:5-25; **Ex. 44** (Grisham Dep. Tr.) 176:22-177:12; **Ex. 43** (Greer Dep. Tr.) 108:16-19; **Ex. 40** (Golden Dep. Tr. (dated Feb. 10, 2026)) 47:21-49:24; **Ex. 60** (Kazi Dep. Tr.) 287:4-24; **Ex. 97** (Rasenberger Dep. Tr. (dated Feb. 12, 2026)) 119:24-121:17.

d) Books Plaintiffs concede they do not expect AI to replace them.

Tucker Decl. ¶¶ 45-46, 68-69; Hurst Decl. **Ex. 25** (Connelly Dep. Tr.) 90:24-91:6; 160:3-13; **Ex. 35** (Franzen Dep. Tr.) 90:21-24; **Ex. 55** (Hwang Dep. Tr.) 152:5-25; **Ex. 44** (Grisham Dep. Tr.) 176:22-177:12; **Ex. 43** (Greer Dep. Tr.) 108:16-19; **Ex. 87** (Picoult Dep. Tr.) 22:10-15; 98:24-99:13; **Ex. 55** (Hwang Dep. Tr.) 150:10–24; **Ex. 9** (Baldacci Dep. Tr.) 231:5–12.

27. No agreements for licensing books as LLM training data existed at the time that the LLMs at issue were trained.

Tucker Decl. ¶¶ 13(a), 90-91; *see also* Hurst Decl. **Ex. 254** (2025.11.21 Hwang

Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 255** (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 257** (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 258** (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 259** (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 293** (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 260** (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 261** (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 253** (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 256** (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.

a) Books Plaintiffs have not offered any evidence that they were approached by any LLM developers to license their works for training;

Hurst Decl. **Ex. 43** (Greer Dep. Tr.) 112:24-113:3; **Ex. 128** (Wolicki Dep. Tr.) 157:16-18, 211:13-16; **Ex. 106** (Schiff Dep. Tr.) 195:19-23; **Ex. 14** (Branch Dep. Tr.) 108:21-109:6; **Ex. 9** (Baldacci Dep. Tr.) 214:16-215:2; **Ex. 87** (Picoult Dep. Tr.) 165:10-15; **Ex. 44** (Grisham Dep. Tr.) 184:17-24; **Ex. 254** (2025.11.21 Hwang Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 255** (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 257** (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 258** (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 259** (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 293** (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; **Ex.**

260 (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 261** (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 253** (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 256** (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.

b) Books Plaintiffs have not offered any evidence that they sought to license their works for LLM training.

Hurst Decl. **Ex. 43** (Greer Dep. Tr.) 112:24-113:3; **Ex. 106** (Schiff Dep. Tr.) 195:19-23; **Ex. 98** (Rasenberger 30(b)(6) Dep. Tr.) 169:2-17; **Ex. 14** (Branch Dep. Tr.) 108:25-109:6; **Ex. 254** (2025.11.21 Hwang Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 255** (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 257** (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 258** (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 259** (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 293** (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 260** (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; **Ex. 261** (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 253** (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; **Ex. 256** (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.

28. There are substantial impediments to the formation of a market for licensing a training corpus sufficient to train the type of general purpose LLMs at issue in this

case. For example:

a) A very large volume and a large variety of textual data is necessary for the training corpus to develop general purpose LLMs like those at issue; as such, no individual work has value by itself as training material nor is it currently technologically possible to isolate the contribution of any particular piece of data to an LLM's output generation;

Lafferty Decl. ¶¶ 16, 34, 94-97, 99, 168, 283, 290, 322, 433; Usdan Decl. ¶¶ 10-11; Hurst Decl. **Ex. 53** (Hoeschele Dep. Tr. (Vol. 1, dated Apr. 2, 2025)) 26:6-22, 35:18-36:11, 16:17-22, 17:1-23, 19:12-20:1, 23:24-24:25 25:15-23; **Ex. 117** (Sutskever Dep. Tr.) 28:5-29:15, 67:9-15, 255:12-256:2.

b) The Asserted Works were a minuscule share of the training corpus used to train the GPT models;

Tucker Decl. ¶¶ 13(b), 95, 99-100; Lafferty Decl. ¶¶ 94-97, 105-109.

c) For a traditional copyright license, rightsholders monetize the inherent value of specific creative expression contained within their individual works. By contrast, when licensing data for LLM training, value is not attributable to any discrete piece of interchangeable data, but rather emerges from statistical distributions observable only across large-scale datasets. There is no settled or workable means for pricing any given piece of training data;

Tucker Decl. ¶¶ 13(c), 111-114.

d) A developer seeking to license a training corpus would need to obtain rights from millions of dispersed authors and content creators;

Tucker Decl. ¶¶ 13(c), 89, 101-106; Usdan Decl. ¶¶ 10-11, 22, 24.

e) Rightsholders for individual works can be difficult and costly to identify and locate,

increasing the transaction costs of negotiating any proposed training data licenses;

Tucker Decl. ¶¶ 13(c), 102-106; Usdan Decl. ¶ 24.

f) Negotiations for the aggregation of the necessary volume and variety of data are susceptible to holdout problems;

Tucker Decl. ¶¶ 13(c), 107-110.

g) When Microsoft licensed book data from [redacted] for LLM training at a price the publisher was willing to accept, the publisher [redacted], and Microsoft ultimately declined to continue that agreement or to pursue any others.

Usdan Decl. ¶¶ 22-25.

29. After the release of the LLMs at issue in this case, Books Plaintiffs have continued to write and publish books.

Tucker Decl. ¶¶ 14, 151; Hurst Decl. **Ex. 128** (Wolicki Dep. Tr.) 26:20-27:3, 30:14-20; **Ex. 87** (Picoult Dep. Tr.) 166:14-167:4; **Ex. 108** (Shapiro Dep. Tr.) 46:1-12; **Ex. 106** (Schiff Dep. Tr.) 20:16-21:6; **Ex. 35** (Franzen Dep. Tr.) 61:14-18; **Ex. 44** (Grisham Dep. Tr.) 178:21-25.

B. Lack Of Volitional Conduct

30. Microsoft did not acquire the Books1 or Books2 datasets used by OpenAI as part of the training corpus in connection with the development of GPT 3 and 3.5.

Hurst Decl. **Ex. 120** (Tinter 30(b)(6) Dep. Tr.) 166:23-169:17, 13:6-22:18, 29:24-30:22; **Ex. 120.A** (Tinter Dep. Ex. 3 - Joint Development and Collaboration Agreement) 4-5 (§§1(f), 1(q)); **Ex. 85** (Parakhin Dep. Tr.) 79:3-18, 15:18-21, 16:1-5, 18:7-12, 22:25-23:8, 23:21-25:6, 26:12-27:4; **Ex. 118** (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 62:4-25, 296:12-18, 13:14-16, 14:14-20, 15:2-12, 16:18-19:10, 22:1-14, 22:25-23:15, 24:2-27:16; **Ex. 96** (Radford Dep. Tr.) 37:17-41:11, 46:3-13, 117:4-20, 11:15-12:10, 13:22-14:6, 16:3-8, 17:13-19, 20:25-24:22; **Ex. 295** (Mann Dep. Tr. (dated Mar. 3, 2025)) 58:14-60:4, 65:22-66:11, 13:19-20:5, 20:21-21:21; **Ex. 48** (Hallacy Dep. Tr. (Vol. 1, dated Sept. 30, 2025)) 91:10-12, 92:5-18, 9:3-13, 9:23-10:9, 11:9-22, 29:16-25, 33:22-34:18; **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 95:10-23, 168:1-11, 169:15-170:17, 9:18-21, 15:2-25, 16:7-11, 17:2-18:7; **Ex. 84** (Paino 30(b)(6) Dep. Tr.) 15:12-21:24, 22:12-26:6, 216:11-217:22, 12:18-25, 13:3-6, 13:22-14:7.

31. Microsoft did not direct OpenAI to download the Books1 or Books2 datasets from Library Genesis ("LibGen") or any other so-called "shadow library."

Hurst Decl. **Ex. 120** (Tinter 30(b)(6) Dep. Tr.) 166:23-169:17; **Ex. 120.A** (Tinter Dep. Ex. 3 – Joint Development and Collaboration Agreement) 4-5 (§§1(f), 1(q)); **Ex. 85** (Parakhin Dep. Tr.) 26:12-27:4, 79:3-18; **Ex. 118** (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 62:4-25, 296:12-18; **Ex. 96** (Radford Dep. Tr.) 37:17-41:11, 46:3-13, 117:4-20; **Ex. 295** (Mann Dep. Tr. (dated Mar. 3, 2025)) 58:14-60:4, 65:22-66:11, 13:19-20:5, 20:21-21:21; **Ex. 48** (Hallacy Dep. Tr. (Vol. 1,

dated Sept. 30, 2025)) 91:10-12, 92:5-18; **Ex. 104** (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 95:10-23, 168:1-11; 169:15-170:17; **Ex. 84** (Paino 30(b)(6) Dep. Tr.) 15:12-21:24, 22:12-26:6, 216:11-217:22.

32. To the extent Microsoft acquired data to be used by OpenAI in connection with the development of LLMs, Microsoft did so by using ordinary software robots that crawled content from URLs across the public internet; Microsoft's web robots were not capable of utilizing the Torrent protocol implemented by so-called "shadow libraries."

Gaggara Decl. ¶¶ 20-21; Hurst Decl. **Ex. 38** (Gaggara 30(b)(6) Dep. Tr. (dated Aug. 20, 2025)) 67:1-20, 143:13-144:6, 8:9-19, 9:25-13:17.

II. COUNT II: VICARIOUS COPYRIGHT INFRINGEMENT

A. No underlying direct infringement

33. Microsoft incorporates Undisputed Facts 1-32 as though fully restated herein.

III. COUNT III: CONTRIBUTORY COPYRIGHT INFRINGEMENT

A. No underlying direct infringement

34. Microsoft incorporates Undisputed Facts 1-33 as though fully restated herein.

Dated: September 4, 2026

Respectfully submitted,

/s/ Annette L. Hurst

ORRICK, HERRINGTON & SUTCLIFFE LLP
Annette L. Hurst (admitted pro hac vice)
Daniel Justice (admitted pro hac vice)
Robert Scott Shwarts
Catherine Lui (admitted pro hac vice)
Amanda H. Schwartz (admitted pro hac vice)
Ana Mendez-Villamil (admitted pro-hac vice)
The Orrick Building
405 Howard Street
San Francisco, CA 94105-2669
Telephone: (415) 773-5700
Facsimile: (415) 773-5759
ahurst@orrick.com
djustice@orrick.com
aschwartz@orrick.com
amendez-villamil@orrick.com

Christopher J. Cariello
Lisa T. Simpson
Marc Shapiro
Vanessa Sorrentino
Hilda Obeng
51 West 52nd Street
New York, NY 10019
Telephone: (212) 506-3778
Facsimile: (212) 506-5151
ccariello@orrick.com
lsimpson@orrick.com
mrshapiro@orrick.com
vsorrentino@orrick.com
hobeng@orrick.com

Sheryl Koval Garko (admitted pro hac vice)
Laura Brooks Najemy (admitted pro hac vice)
222 Berkley Street
Boston, MA 02116
Telephone: (617) 880-1800
Facsimile: (617) 880-1801
sgarko@orrick.com
lnajemy@orrick.com

Isaac Sawlih Behnawa (admitted pro hac vice)
Johnathan Vaknin (admitted pro hac vice)
355 S. Grand Ave. Suite 2700
Los Angeles, CA 90071
Telephone: (213) 612-2179
Facsimile: (213) 612-2499
ibehnawa@orrick.com
jvaknin@orrick.com

Monica Svetoslavov (admitted pro hac vice)
Louis Andrew Schreiber
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 339-8400
Facsimile: (202) 339-8500
msvetoslavov@orrick.com
aschreiber@orrick.com

FAEGRE DRINKER BIDDLE & REATH LLP
Jeffrey S. Jacobson
1177 Avenue of the Americas
New York, New York 10036
Telephone: (212) 248-3191
jeffrey.jacobson@faegredrinker.com

Jared B. Briant (admitted pro hac vice)
Kirstin L. Stoll-DeBell (admitted pro hac vice)
Timothy Scull (admitted pro hac vice)
Shelby Pickar-Dennis (admitted pro hac vice)
1144 15th Street, Suite 3400
Denver, Colorado 80202
Telephone: (303) 607-3588
jared.briant@faegredrinker.com
kirstin.stolldebell@faegredrinker.com
timothy.scull@faegredrinker.com
shelby.pickar-dennis@faegredrinker.com

Lora A. Brzezynski (admitted pro hac vice)
Brianna Lynn Silverstein (admitted pro hac vice)
Faith Maggard (admitted pro hac vice)
1500 K Street NW, Suite 1100
Washington, DC 20005
Telephone: (202) 230-5000
Facsimile: (202) 842-8465
lora.brzezynski@faegredrinker.com

brianna.silverstein@faegredrinker.com
faith.maggard@faegredrinker.com

Carrie A. Beyer (admitted pro hac vice)
320 South Canal Street, Suite 3300
Chicago, IL 60606-5707
Telephone: (312) 569-1000
Facsimile: (312) 569-3000
carrie.beyer@faegredrinker.com

Elizabeth M.C. Scheibel (admitted pro hac vice)
2200 Wells Fargo Center, 90 S. 7th Street
Minneapolis, MN 55402
Telephone: (612) 766-7000
Facsimile: (612) 766-1600
elizabeth.scheibel@faegredrinker.com

Andrew Lawrence Van Houter
600 Campus Drive
Florham Park, NJ 07932
Telephone: (973) 549-7018
Facsimile: (973) 360-9831
Andrew.vanhouter@faegredrinker.com

Attorneys for Defendant Microsoft Corporation