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-08292-SHS-OTW Case No. 1:23-cv-10211-SHS-OTW	**ECF CASE** 1:25-MD-03143-SHS/OTW

CLASS PLAINTIFFS' CORRECTED RULE 56.1 STATEMENT OF UNDISPUTED MATERIAL FACTS IN SUPPORT OF CLASS PLAINTIFFS' MOTION FOR PARTIAL SUMMARY JUDGMENT

TABLE OF CONTENTS

I. The Parties, Asserted Works, and Accused Models 1

A. Class Plaintiffs and Asserted Works 1

1. David Baldacci and Columbus Rose, Ltd. 2

2. Taylor Branch 2

3. Michael Connelly and Hieronymus Inc. 3

4. Sylvia Day and Sylvia Day LLC 3

5. Jonathan Franzen 4

6. Christopher Golden and Daring Greatly Corporation 4

7. Andrew Sean Greer 4

8. John Grisham and Belfry Holdings, Inc. 5

9. David Henry Hwang 5

10. George R.R. Martin & Wo & Shade, LLC 6

11. Jodi Picoult 6

12. Stacy Schiff and Gattopardo, LLC 7

13. James Shapiro 7

14. The Authors Guild (Estate of Mignon Eberhart) 8

B. The OpenAI Entities 8

C. OpenAI's Conversion from Nonprofit to Commercial Enterprise 10

D. Generative AI and Large Language Models 11

E. The Accused Models and ChatGPT 13

F. Microsoft 15

II. OpenAI Downloaded and Stored Millions of Pirated Books from Libgen It Could Have Bought 16

A. Library Genesis 16

B. July–November 2018: OpenAI Priced Lawful Book Acquisition and Chose LibGen Instead 18

C. November 2018–February 2019: OpenAI Budgeted to Buy Books, Then Built Its First LibGen Dataset Instead 21

D. February–July 2019: OpenAI Presented LibGen to Microsoft and Closed a $1 Billion Investment 24

E. September 2019–January 2020: Mann and Hallacy Torrented the Entire LibGen Collection 25

F. 2019–June 2022: OpenAI Stored the Pirated Books in Locations Accessible to Its Employees and Put Them to Continued Use 33

1. Where OpenAI Stored the Books 33
2. Who Could Access the Books 35
3. Copies, Migration, and Retention 36
4. Continued Use of the LibGen Books 37
G. April–May 2020: OpenAI Renamed LibGen "Books1" and "Books2" 38
H. April 2021–2024: OpenAI Repeatedly Priced Lawful Alternatives and Never Replaced the LibGen Books 41
I. December 2021–Summer 2022: OpenAI Deleted LibGen 45
J. Throughout, OpenAI Knew LibGen Was "Sketchy [as F---]" 49
1. OpenAI's Contemporaneous Knowledge 49
2. Defendants' Admissions That Piracy Is Wrong and Illegal 57
III. OpenAI Also Downloaded Books from Books3, Targeted Scrapes, and Web Crawls 61
A. Books3 and The Pile 61
B. OpenAI Deleted Its Copy of The Pile in October 2024 66
C. OpenAI Also Acquired Books Through Targeted Scrapes and Third-Party Web Crawls 67
IV. OpenAI Trained Its LLMs on Datasets Containing Books 71
A. OpenAI Knew Books Were High-Quality Training Data 71
B. Defendants' Experts Acknowledge That Books Were Not Necessary 79
C. OpenAI Used LibGen to Train GPT-3 and GPT 3.5 83
D. OpenAI Also Wanted to Use Books3 to Train Its Models 85
E. OpenAI Also Trained the Accused Models on Datasets Derived from Its Scrapes and Crawls 85
F. OpenAI Created Multiple Copies of Books in Training 86
G. The Accused Models Reproduce Class Plaintiffs' Works to End Users 87
V. OpenAI and Microsoft Distributed Books to Each Other and Third Parties 91
A. OpenAI Provided Microsoft Access to GPT-3 Training Data 91
B. Microsoft Transferred Crawled Data to OpenAI (Project Mango) 95
C. OpenAI Transferred LibGen to Book-Summarization Contractors 97
VI. Microsoft's Role in OpenAI's Infringement 98
A. OpenAI and Microsoft Raced to Close Google's Lead in Compute, Talent, and Books 98
B. Microsoft Formed a Partnership with OpenAI 101
C. Microsoft Knew OpenAI Used LibGen 105

D. Microsoft Built Bespoke Supercomputing Systems for OpenAI's Model Training. 109

E. Microsoft Encouraged OpenAI's Data Practices, [redacted], and Obtained OpenAI's Technology 111

F. Microsoft Could Control OpenAI's Training Data 113

G. Microsoft Benefitted Financially from OpenAI's Infringement 116

VII. The Licensing Market and Market Harm 117

A. Defendants Recognized the Value of Books and the Harm to Authors 117

B. Defendants' Book-Licensing Agreements 129

C. Other Licensing Activities 137

D. The Proliferation of AI-Generated Books 141

E. AI Poses Substantial Risks That Defendants Recognized 150

VIII. Copies of Plaintiffs' Works Are Present in OpenAI's Data 153

A. Pirated Books 155

B. Training Datasets 156

Pursuant to Local Rule 56.1, Plaintiffs David Baldacci, Taylor Branch, Michael Connelly, Sylvia Day, Jonathan Franzen, Christopher Golden, Andrew Sean Greer, John Grisham, David Henry Hwang, Matthew Klam, George R.R. Martin, Jodi Picoult, Stacy Schiff, James Shapiro, The Authors Guild, Belfry Holdings, Inc., Columbus Rose, Ltd., Daring Greatly Corporation, Gattopardo, LLC, Hieronymus, Inc., Sylvia Day LLC, and Wo & Shade LLC (collectively, "Class Plaintiffs") respectfully submit this Statement of Undisputed Material Facts in Support of Class Plaintiffs' Motion for Partial Summary Judgment.

I. The Parties, Asserted Works, and Accused Models

A. Class Plaintiffs and Asserted Works

1. Class Plaintiffs are the bestselling and award-winning book authors identified below, the entities through which certain of those authors conduct their professional activities, and The Authors Guild, the oldest and largest professional organization for published writers in the United States.

2. The works at issue in this Statement are the copyrighted books listed in Appendix A to this Statement (the "Asserted Works"), which sets out the author, copyright holder, copyright registration number, first publication date, effective copyright registration date, and publishing royalty recipient for each work. While Plaintiffs' motion addresses the Asserted Works, Plaintiffs proudly represent the interests of copyright holders who were harmed by Defendants' conduct. Nelson Decl. Ex. 51 at 186:14-187:19; Nelson Decl. Ex. 5 at 220:7-221:14.

3. The copyright claimant registered each Asserted Work with the United States Copyright Office within five years of its first publication. Appx. A.

4. Each Asserted Work is currently available for sale. Appx. A.

1. David Baldacci and Columbus Rose, Ltd.

5. Plaintiff David Baldacci is an author, philanthropist, and lawyer who resides in Virginia. Nelson Decl. Ex. 104.

6. Baldacci has published more than fifty novels for adults, every one a national and international bestseller, and his books have sold more than 200 million copies in over eighty countries. *Id.*

7. Plaintiff Columbus Rose, Ltd. is a company through which David Baldacci conducts his professional activities, including the ownership and management of copyrights in his literary works. Nelson Decl. Ex. 5 at 104:6–19.

8. Columbus Rose, Ltd. owns the valid copyrights in forty published books authored by David Baldacci, including titles in the *Camel Club* series, the *King & Maxwell* series, the *Will Robie* series, and the *Amos Decker* series. Appx. A.

2. Taylor Branch

9. Plaintiff Taylor Branch is an author and historian who resides in Maryland. Nelson Decl. Ex. 69.

10. Branch's *Parting the Waters* won the Pulitzer Prize for History and the National Book Critics Circle Award, and it opens Branch's three-volume *America in the King Years* chronicle of Martin Luther King, Jr. and the civil rights movement. *Id.*

11. Branch has also received a MacArthur Fellowship and the National Humanities Medal. *Id.*

12. Taylor Branch owns the valid copyrights in two published books: *Parting the Waters: America in the King Years, 1954-63* and *Pillar of Fire: America in the King Years, 1963-65*. Appx. A.

3. Michael Connelly and Hieronymus Inc.

13. Plaintiff Michael Connelly is an author who resides in Florida. Nelson Decl. Ex. 110.

14. Connelly has written more than forty novels that have sold over ninety million copies worldwide, and his debut, *The Black Echo*, won the Mystery Writers of America Edgar Award for Best First Novel. *Id.*

15. He executive produces *Bosch*, *Bosch: Legacy*, and *Ballard* on Amazon Prime Video and *The Lincoln Lawyer* on Netflix, all of which are based on his books. *Id.*

16. Plaintiff Hieronymus Inc. is a corporation through which Michael Connelly conducts his professional activities, including the ownership and management of copyrights in his literary works. Nelson Decl. Ex. 15 at 33:8–13.

17. Hieronymus Inc. owns the valid copyrights in thirty-six published books authored by Michael Connelly, including the *Harry Bosch* series, the *Mickey Haller* (Lincoln Lawyer) series, and the *Renee Ballard* series. Appx. A.

4. Sylvia Day and Sylvia Day LLC

18. Plaintiff Sylvia Day is an author who resides in Nevada. Nelson Decl. Ex. 124.

19. Day has written more than twenty award-winning novels translated into forty-one languages and is a number one bestseller in twenty-nine countries. *Id.*

20. Plaintiff Sylvia Day LLC is a limited liability company through which Sylvia Day conducts her professional activities, including the ownership and management of copyrights in her literary works. Nelson Decl. Ex. 66 at 31:1–11; Nelson Decl. Ex. 137.

21. Sylvia Day LLC and Sylvia Day own the valid copyrights in twenty-nine published books authored by Sylvia Day, including the *Crossfire* series, which begins with *Bared to You*. Appx. A.

5. Jonathan Franzen

22. Plaintiff Jonathan Franzen is an author who resides in California. Nelson Decl. Ex. 100.

23. Franzen's *The Corrections* won the National Book Award and was ranked on the New York Times Book Review's 100 Best Books of the 21st Century. Nelson Decl. Ex. 101.

24. *Time* magazine featured Franzen on its cover as the "Great American Novelist," with an article marking the publication of his novel, Freedom. Nelson Decl. Ex. 97.

25. Jonathan Franzen owns the valid copyrights in five published books, including *The Corrections*, *Freedom*, and *Crossroads*. Appx. A.

6. Christopher Golden and Daring Greatly Corporation

26. Plaintiff Christopher Golden is an author who resides in Massachusetts. Nelson Decl. Ex. 103.

27. Golden is a *New York Times* bestselling horror and thriller novelist who has been called the "the king of horror-thriller." *Id.*

28. His work has been published in dozens of languages, and he is the winner of the Bram Stoker Award. *Id.*

29. Plaintiff Daring Greatly Corporation is a Massachusetts-based corporation through which Christopher Golden conducts his professional activities, including the management of royalties in his literary works. Nelson Decl. Ex. 717 at 257:3–11.

30. Christopher Golden owns the valid copyright in one published book: *Ararat*. Appx. A.

7. Andrew Sean Greer

31. Plaintiff Andrew Sean Greer is an author who resides in California. Nelson Decl. Ex. 70.

32. Greer's *Less* won the 2018 Pulitzer Prize for Fiction and was a national bestseller. *Id.*

33. Greer has also received a Guggenheim Fellowship, the California Book Award, and the New York Public Library Young Lions Award. *Id.*

34. Andrew Sean Greer owns the valid copyrights in four published books, including *Less* and its sequel *Less Is Lost*, as well as *The Confessions of Max Tivoli* and *The Story of a Marriage*. Appx. A.

8. <u>John Grisham and Belfry Holdings, Inc.</u>

35. Plaintiff John Grisham is an author who resides in Virginia. Nelson Decl. Ex. 85.

36. Grisham has published more than fifty consecutive number one bestsellers, and his books have been translated into nearly fifty languages. *Id.*

37. He has twice won the Harper Lee Prize for Legal Fiction and received the Library of Congress Creative Achievement Award for Fiction. *Id.*

38. Plaintiff Belfry Holdings, Inc. is a corporation through which John Grisham conducts his professional activities, including the ownership and management of copyrights in his literary works. Nelson Decl. Ex. 24 at 16:10–25.

39. Belfry Holdings, Inc. and John Grisham own the valid copyrights in twenty-three published books authored by John Grisham, including *A Time to Kill*, *The Firm*, *The Pelican Brief*, and *The Client*. Appx. A.

9. <u>David Henry Hwang</u>

40. Plaintiff David Henry Hwang is an author and playwright who resides in New York. Nelson Decl. Ex. 793.

41. He has been described by the New York Times as "a true original" and by *Time* magazine as "the first important dramatist of American public life since Arthur Miller." Nelson Decl. Ex. 122.

42. Hwang's *M. Butterfly* won the 1988 Tony Award for Best Play, making Hwang the first Asian American playwright to win that award, and was a Pulitzer Prize finalist. Nelson Decl. Ex. 94.

43. David Henry Hwang owns the valid copyrights in three published books, including *M. Butterfly*. Appx. A.

10. <u>George R.R. Martin & Wo & Shade, LLC</u>

44. Plaintiff George R. R. Martin is an author who resides in New Mexico. Nelson Decl. Ex. 83. Martin's *A Song of Ice and Fire* novels have sold more than 85 million copies worldwide and appear in 47 languages. Nelson Decl. Ex. 117.

45. He has won numerous Hugo and Nebula Awards, and the World Fantasy Convention honored him with its Life Achievement Award in 2012. Nelson Decl. Ex. 107.

46. HBO adapted *A Song of Ice and Fire* as *Game of Thrones*, which earned Martin three Emmy Awards as a producer. *Id.*

47. Plaintiff Wo & Shade, LLC is a limited liability company through which George R. R. Martin conducts his professional activities, including the ownership and management of copyrights in his literary works. Nelson Decl. Ex. 34 at 16:13–19.

48. Wo & Shade, LLC owns the valid copyrights in fifteen published books authored by George R.R. Martin, including novels in the *A Song of Ice and Fire* series. Appx. A.

11. <u>Jodi Picoult</u>

49. Plaintiff Jodi Picoult is an author who resides in New Hampshire. Nelson Decl. Ex. 86.

50. Picoult has written twenty-nine novels translated into forty languages, with an estimated 63.7 million copies in print. *Id.*

51. Her honors include the New England Bookseller Award for Fiction and the Sarah Josepha Hale Award. *Id.*

52. Jodi Picoult owns the valid copyrights in twenty-six published books, including *My Sister's Keeper*, *Nineteen Minutes*, and *Small Great Things*. Appx. A.

12. <u>Stacy Schiff and Gattopardo, LLC</u>

53. Plaintiff Stacy Schiff is an author who resides in New York. Nelson Decl. Ex. 95.

54. She is a Pulitzer Prize winner, and her later books, among them *Cleopatra: A Life* and *The Revolutionary: Samuel Adams*, have been national bestsellers. *Id.*

55. Schiff has also received fellowships from the Guggenheim Foundation and the National Endowment for the Humanities and was a Director's Fellow at the Cullman Center for Scholars and Writers at the New York Public Library. *Id.*

56. Plaintiff Gattopardo, LLC is a limited liability company through which Stacy Schiff conducts her professional activities, including the management of royalties in her literary works. Nelson Decl. Ex. 51 at 16:7–12.

57. Stacy Schiff owns the valid copyrights in three published books, including *The Witches* and *Cleopatra: A Life*. Appx. A.

13. <u>James Shapiro</u>

58. Plaintiff James Shapiro is an author who resides in New York. Nelson Decl. Ex. 82.

59. Shapiro is the Larry Miller Professor Emeritus of English and Comparative Literature at Columbia University and the Shakespeare Scholar in Residence at the Public Theater in New York. Nelson Decl. Ex. 82; Nelson Decl. Ex. 109.

60. *A Year in the Life of William Shakespeare: 1599* won the Samuel Johnson Prize, and *The Year of Lear* won the James Tait Black Memorial Prize. Nelson Decl. Ex. 109.

61. James Shapiro owns the valid copyrights in three published books, including *A Year in the Life of William Shakespeare: 1599* and *The Year of Lear: Shakespeare in 1606*. Appx. A.

14. <u>The Authors Guild (Estate of Mignon Eberhart)</u>

62. Plaintiff the Authors Guild is a registered 501(c)(6) not-for-profit association that advocates for professional writers and provides services to more than 15,000 members. Nelson Decl. Ex. 46(a) at 21:13-22:9.

63. Founded in 1912, the Authors Guild is the oldest and largest professional organization for published writers in the United States. Nelson Decl. Ex. 73.

64. Mignon Eberhart was an author who published more than fifty mystery novels over a sixty-year career. Nelson Decl. Ex. 87.

65. She is a former president of the Mystery Writers of America and a recipient of the organization's Grand Master Award. *Id.*

66. Eberhart bequeathed the copyrights in her published books to the Authors Guild, which acquired those copyrights upon her death. Nelson Decl. Ex. 136.

67. The Authors Guild owns the valid copyrights in five published books authored by Mignon Eberhart. Appx. A.

B. <u>The OpenAI Entities</u>

68. Defendant OpenAI, Inc. (now known as OpenAI Foundation) is a Delaware corporation headquartered in California. OpenAI's Answer ¶ 44, ECF No. 754.

69. Before 2019, OpenAI, Inc. was involved in designing and developing OpenAI's GPT-based products and services. Nelson Decl. Ex. 30 at 115:15–19.

70. Defendant OpenAI GP, LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 45, ECF No. 754.

71. During at least some portion of the relevant time period, OpenAI, Inc. was the sole member of OpenAI GP, LLC. Nelson Decl. Ex. 704.

72. Defendant OpenAI OpCo LLC (formerly known as OpenAI LP) is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 46, ECF No. 754.

73. Since 2019, OpenAI OpCo LLC has been involved in designing and developing OpenAI's GPT-based products and services. Nelson Decl. Ex. 30 at 115:8–23.

74. Defendant OpenAI, LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 47, ECF No. 754.

75. During at least some portion of the relevant time period, OpenAI, LLC distributed OpenAI's products. Nelson Decl. Ex. 30 at 62:2–62:16.

76. Defendant OpenAI Global LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 48, ECF No. 754.

77. During at least some portion of the relevant time period, Microsoft Corporation held a minority stake in OpenAI Global LLC. Nelson Decl. Ex. 704.

78. Defendant OpenAI Holdings, LLC is a Delaware limited liability company. OpenAI's Answer ¶ 50, ECF No. 754.

79. During at least some portion of the relevant time period, OpenAI Holdings, LLC was the majority member in OpenAI Global LLC. Nelson Decl. Ex. 704.

80. On March 4, 2026, OpenAI moved to substitute OpenAI Group PBC for OpenAI Holdings, LLC because "on December 31, 2025, OpenAI Holdings, LLC merged into OpenAI

Group PBC and no longer exists as a standalone entity." ECF No. 1386 at 2. The Court granted the motion on March 26, 2026. ECF No. 1464 at 1.

81. Defendant OpenAI Startup Fund I LP is a Delaware limited partnership headquartered in California. Consolidated Amended Complaint ¶ 51, ECF No. 138; OpenAI's Answer ¶ 51, ECF No. 754.

82. Defendant OpenAI Startup Fund GP I LLC is a Delaware limited liability company headquartered in California. Consolidated Amended Complaint ¶ 51, ECF No. 138; OpenAI's Answer ¶ 52, ECF No. 754.

83. Defendant OpenAI Startup Fund Management LLC is a Delaware limited liability company headquartered in San Francisco, California. Consolidated Amended Complaint ¶ 51, ECF No. 138; OpenAI's Answer ¶ 53, ECF No. 754.

84. This Statement collectively refers to OpenAI, Inc., OpenAI GP, LLC, OpenAI, LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC and its successor by merger, OpenAI Group PBC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, and OpenAI Startup Fund Management LLC as "OpenAI," and refers to OpenAI and Microsoft Corporation collectively as "Defendants."

C. <u>OpenAI's Conversion from Nonprofit to Commercial Enterprise</u>

85. OpenAI began in December 2015 as a nonprofit research organization. Nelson Decl. Ex. 825. At that time, OpenAI told the Internal Revenue Service that it did "not plan to play any role in developing commercial products or equipment" and "intend[ed] to make its research freely available to the public." Nelson Decl. Ex. 705 at -963.

86. That structure did not hold. By 2017, OpenAI's founders were considering an organizational change that would allow investment rather than donations alone. Altman testified

that "it became clear that we were going to need far more capital than we thought we were going to need when we started the project." Nelson Decl. Ex. 2(a) at 171:8–21.

87. In March 2019, OpenAI formed OpenAI LP, a for-profit subsidiary it called a "capped-profit" company. Nelson Decl. Ex. 706. OpenAI LP could issue equity to raise capital and took over the majority of OpenAI's operations. *Id.*

88. OpenAI co-founder Greg Brockman recorded the reasons for that change in his contemporaneous journal entries, writing that "we've been thinking that maybe we should just flip to a for profit" because "making the money for us sounds great," and that "we just want something that'll get us the money." Nelson Decl. Ex. 365 at -557-558.

89. Brockman also wrote that he "was deeply motivated by the gazillions." Nelson Decl. Ex. 367 at -624.

90. In 2019, Brockman expressed surprise that OpenAI "actually got the for-profit [they] were dreaming of," and identified his personal goals as "money + fame/respect." Nelson Decl. Ex. 366.

91. OpenAI's valuation has since reached at least $852 billion. Nelson Decl. Ex. 707.

92. In July 2026, Reuters reported that OpenAI is targeting a valuation of up to $1 trillion in a stock market debut that could come as early as September. Nelson Decl. Ex. 708.

93. The products OpenAI developed and commercialized through that for-profit structure are the GPT large language models and ChatGPT described below. *See infra* Section I.E.

D. <u>Generative AI and Large Language Models</u>

94. Large language models ("LLMs") are a kind of "generative artificial intelligence," a term that refers to systems that generate outputs, such as text and images, based on the inputs (i.e. data) on which the system was trained. Nelson Decl. Ex. 679.

95. The core of the LLM training process is called "pre-training." All of the in-suit models went through this "pre-training" process. Nelson Decl. Ex. 680.

96. During pre-training, LLMs are trained using a process by which the in-development model is exposed to incomplete text data, such as part of a sentence, drawn from a corpus of "training data." Nelson Decl. Ex. 681. The model then outputs a prediction of the next word—or fragment of a word, known as a "token"— in the sequence. *Id*. At the start of this pre-training process, outputs are effectively random. *Id*.

97. The model's predicted output is then checked against the training data. If the model's output does not match the training data, the model training code adjusts the model's mathematical components so that the incorrect guess is penalized and the correct answer is favored. Nelson Decl. Ex. 700.

98. These internal components are known as "weights" or "parameters." The measurement of the extent to which the model's prediction varies from the actual data in its training data is known as "loss." Nelson Decl. Ex. 700.

99. OpenAI trained each of the Accused Models using the "next-token prediction loss" method. Nelson Decl. Ex. 50(a) at 175:24–176:1 ("For all of the models we are talking about, they were trained with next-token prediction loss.").

100. Under that method, pre-training seeks to minimize the loss—that is, to train the model so that its outputs match the training data as closely as possible. Nelson Decl. Ex. 50(a) at 175:6–21, 178:9–14.

101. The core function of an LLM is to generate text that is similar to the dataset of text on which it was trained. Nelson Decl. Ex. 298 at -468 (internal OpenAI presentation describing

how "the training objective" makes the "models want to regurgitate" the text on which the model was trained).

102. James Betker, an OpenAI employee who has "trained a lot of generative AI models," stated that LLMs are "truly approximating their datasets to an incredible degree." Nelson Decl. Ex. 709.

103. As he put it, "model behavior is not determined by [technical choices around model training such as] architecture, hyperparameters, or optimizer choices. It's determined by your dataset, nothing else. Everything else is a means to an end in efficiently deliver[ing] compute to approximating that dataset." *Id.* Although companies give their models names such as "'Lambda,' 'ChatGPT,' 'Bard,' or 'Claude,' . . . it's not the model weights that you are referring to. It's the dataset." *Id.*

E. <u>The Accused Models and ChatGPT</u>

104. The models at issue are GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4o mini (the "Accused Models"). Opinion & Order (Oct. 27, 2025), ECF 707 at 1.

105. OpenAI publicly described GPT-3, a 175-billion-parameter model, in a May 2020 paper titled "Language Models are Few-Shot Learners," and released it in June 2020 through an application programming interface. Nelson Decl. Ex. 162 at -653.

106. That interface was OpenAI's first commercial product, and OpenAI described it as a "revenue source to help [the company] cover costs in pursuit of [its] mission." Nelson Decl. Ex. 720.

107. OpenAI finished training GPT-3.5 in early 2022 and released ChatGPT to the public on November 30, 2022, as a free "research preview." Nelson Decl. Ex. 718.

108. ChatGPT is a conversational interface that OpenAI "fine-tuned from a model in the GPT-3.5 series," and OpenAI trained both ChatGPT and GPT-3.5 on Microsoft's Azure supercomputing infrastructure. Nelson Decl. Ex. 718; Nelson Decl. Ex. 36 at 40:5-15.

109. ChatGPT reached 100 million monthly active users within two months—the fastest-growing consumer internet application in history. Nelson Decl. Ex. 721.

110. In February 2023, OpenAI launched ChatGPT Plus, a $20-per-month subscription initially powered by GPT-3.5. OpenAI followed with ChatGPT Enterprise in August 2023 and ChatGPT Team in January 2024. Lasinski Decl. Ex. A at ¶ 43.

111. OpenAI released GPT-3.5 Turbo on March 1, 2023, GPT-4 on March 14, 2023, GPT-4 Turbo in November 2023, GPT-4o in May 2024, and GPT-4o mini in July 2024, and it made each of those models available to users through ChatGPT and through its application programming interface. Lasinski Decl. Ex. A at ¶¶ 41, 43.

112. OpenAI trained GPT-4 in late 2021 and 2022 and released it on March 14, 2023. Lasinski Decl. Ex. A at ¶ 41.

113. OpenAI retired GPT-4 in April 2025 and removed GPT-4o from ChatGPT in February 2026. Lasinski Decl. Ex. A at ¶ 42.

114. ChatGPT reached more than 100 million weekly active users within a year of launch and approximately 350 million within two years. Nelson Decl. Ex. 725 at 10.

115. By July 2025, ChatGPT had 700 million weekly active users—roughly ten percent of the world's adult population—who were sending more than 18 billion messages each week, and OpenAI reported more than 900 million weekly active users by December 2025. Nelson Decl. Ex. 725 at 1; Lasinski Decl. Ex. A at ¶ 77.

116. Between 2019 and 2024, OpenAI generated approximately [redacted] in revenue relating to the Accused Models and the products through which it offered them. Nelson Decl. Ex. 353.

117. OpenAI's revenue grew from [redacted] in 2023 to $3.7 billion in 2024, and OpenAI forecast $12.7 billion in revenue for 2025. Nelson Decl. Ex. 326 at -677. At the gross profit level OpenAI [redacted] in 2023, [redacted] in 2024, and an estimated [redacted] in 2025. *Id.*

118. ChatGPT and the OpenAI application programming interface [redacted] OpenAI's revenues and gross profits in 2023 and 2024, and for [redacted] in 2025. Nelson Decl. Ex. 326 at -677.

119. OpenAI derived nearly [redacted] from paid versions of ChatGPT between 2023 and 2024, and [redacted] from its application programming interface between 2020 and 2024, of which [redacted] is attributable to the Accused Models. Nelson Decl. Ex. 353.

120. OpenAI also received approximately [redacted] from Microsoft under the parties' [redacted]. Nelson Decl. Ex. 213.

F. <u>Microsoft</u>

121. Defendant Microsoft Corporation is a Washington-based corporation that develops and sells software, cloud computing services, and artificial intelligence products. Microsoft's Azure platform supplied the supercomputing infrastructure on which OpenAI trained and deployed its GPT models from GPT-3 forward, in addition to providing storage for Microsoft's training data, which OpenAI understood to be [redacted] Nelson Decl. Ex. 36 at 41:4–11; 60:23–63:6; 85:25–87:2.

122. In July 2019, Microsoft and OpenAI entered a strategic partnership documented in the Joint Development and Collaboration Agreement and related agreements, under which Microsoft committed $1 billion in capital, became OpenAI's exclusive cloud provider, and obtained rights to commercialize OpenAI's technology. Nelson Decl. Ex. 145; Nelson Decl. Ex. 203.

123. Microsoft's total capital commitment to OpenAI has reached $13 billion. Nelson Decl. Ex. 150; Nelson Decl. Ex. 146 at -592; Nelson Decl. Ex. 203; *see infra* § VI.B.

124. In September 2020, Microsoft obtained an exclusive license to GPT-3. Lasinski Decl. Ex. A at ¶ 73.

125. Microsoft receives 20 percent of OpenAI's revenue, held an approximately [redacted] equity stake in OpenAI Group PBC as of September 2025, and has deployed OpenAI's GPT models across more than [redacted] Microsoft products, including Bing, Microsoft Copilot, and Microsoft 365 Copilot. Nelson Decl. Ex. 146 at -596; Nelson Decl. Ex. 40 at 131:21–132:11; Nelson Decl. Ex. 736 at 27; Thornburgh Decl. (ECF 585) ¶ 8; Thornburgh Decl. (ECF 1088) ¶ 8.

II. OpenAI Downloaded and Stored Millions of Pirated Books from Libgen It Could Have Bought

A. Library Genesis

126. Library Genesis ("LibGen") is a shadow library—a website that collects, hosts, and distributes digital copies of copyrighted books without charge and without authorization from the authors and publishers who hold the rights to those books. Seeley Decl. Ex. A at ¶¶ 15–16; Nelson Decl. Ex. 13 at 19:20–20:20; Nelson Decl. Ex. 64 at 22:14–23:12.

127. LibGen originated in Russia in 2008. Seeley Decl. Ex. A at ¶ 21. At its inception, it incorporated copyrighted books from earlier pirate libraries. *Id.*

128. Defendants' own witnesses understood that LibGen distributes large collections of unauthorized copies of copyrighted works, including books. Nelson Decl. Ex. 13 at 19:20–20:20, 21:1–4; Nelson Decl. Ex. 64 at 22:14–23:12.

129. The only person ever to submit a declaration in court on behalf of LibGen "refer[red] to the websites' activities as 'pirating.'" Nelson Decl. Ex. 726.

130. Courts have repeatedly ordered LibGen to shut down and have entered substantial judgments against it. In late October 2015, the District Court for the Southern District of New York ordered LibGen to shut down and to suspend use of the domain name libgen.org. *See id.* at *4–6. The site remained accessible through alternate domains. Seeley Decl. Ex. A at ¶ 21.

131. In 2017, Elsevier obtained a $15 million judgment and a permanent injunction against LibGen. *See* Nelson Decl. Ex. 808; Seeley Decl. Ex. A at ¶ 21.

132. In 2023, a group of United States educational publishers sued LibGen for infringement. Nelson Decl. Ex. 732 at 1. The publishers obtained a default judgment in 2024 for injunctive relief and $30 million. *See id.* at 5–6; Seeley Decl. Ex. A at ¶ 21.

133. The Office of the United States Trade Representative has repeatedly identified LibGen as a notorious market for counterfeiting and piracy. USTR first listed LibGen in its 2017 Out-of-Cycle Review of Notorious Markets. Seeley Decl. Ex. A at ¶ 24. USTR listed LibGen again in its 2019, 2020, 2021, 2023, 2024, and 2025 Reviews of Notorious Markets for Counterfeiting and Piracy. *Id.* at ¶ 44; Shan Decl. Ex. A. at ¶ 46; Nelson Decl. Ex. 728.

134. The 2023 Review reports that LibGen hosts 2.4 million non-fiction books and 2.2 million fiction books, "many of which are unauthorized copies of copyright protected content." Nelson Decl. Ex. 729.

135. United States law enforcement has also acted against LibGen and other shadow libraries. *See, e.g.*, Seeley Decl. Ex. A at ¶ 21; Nelson Decl. Ex. 731. In 2022, United States law enforcement seized hundreds of mirror site domains associated with LibGen. Seeley Decl. Ex. A at ¶ 21; Nelson Decl. Ex. 733. LibGen and its network of mirror sites remain active. Seeley Decl. Ex. A at ¶ 21.

136. LibGen divides its books collection into two categories, fiction and nonfiction. Shan Decl. Ex. A. at ¶ 47. LibGen assigns every uploaded file an Md5 hash, a unique code that identifies that file. *Id.* LibGen lists metadata for the files it collects, including title, author, ISBN, upload date, file format, and Md5 hash. *Id.*

137. As of October 2019, LibGen fiction contained 2.18 million books. *Id.* at ¶ 47. As of January 2020, LibGen nonfiction contained 2.46 million books. *Id.* During the period relevant to this action, LibGen hosted approximately 4.6 million books.

B. <u>July–November 2018: OpenAI Priced Lawful Book Acquisition and Chose LibGen Instead</u>

138. In July 2018, OpenAI wrote a document titled "Motivation," to which Alec Radford and Daniela Amodei contributed comments. Nelson Decl. Ex. 303.

139. The document identified "Books," and specifically "Lib Genesis," i.e. Library Genesis, as one source that could help OpenAI "explore how far this approach can go." *Id.*

140. In September 2018, Radford asked in an OpenAI Slack channel whether anyone had links to high-quality text. Nelson Decl. Ex. 243. Miles Brundage replied that recent books were under-represented. *Id.* Radford responded that "modern books are very under-represented due to copyright." *Id.* at -899.

141. In October 2018, OpenAI received an offer to buy [redacted] Nelson Decl. Ex. 332 at -638. Dario Amodei responded that "[redacted]

[redacted] *Id.* at -637.

142. Radford disagreed, replying in the same thread: "Given the scope and scale of book digitization projects, I imagine this would be a [redacted] *Id.*

143. A few days later, Daniela Amodei followed up in the thread: "Coming back to this – [redacted] That works for me if so[.]" *Id.*

144. Radford responded: "That's my current leaning, yep!" *Id.*

145. OpenAI did not purchase or license books in 2018. Nelson Decl. Ex. 45 at 147:13–21. Radford instead downloaded books from LibGen at no cost. *Id.* at 24:19–22, 147:23–148:13.

146. On October 10, 2018, Radford wrote a custom script to download books from LibGen. Nelson Decl. Ex. 691; Nelson Decl. Ex. 45 at 24:19–22, 49:23–50:8, 64:15–65:23, 66:13–67:21, 92:6–94:6.

147. The script downloaded only .epub and .mobi files, which Radford identified using LibGen's metadata. Nelson Decl. Ex. 45 at 49:23–50:8, 64:15–65:23; 67:1–67:16; 92:6–94:6.

148. Radford performed a direct download over HTTP from a LibGen mirror rather than torrenting the files. *Id.* at 43:19–21; Choffnes Decl. Ex. A at ¶¶ 149–51.

149. The script ran for approximately one month and downloaded approximately 117,859 books. Nelson Decl. Ex. 45 at 24:1–18; Shan Decl. Ex. A. at ¶ 49; Choffnes Decl. Ex. A at ¶ 148.

150. Radford ran the download from a personal computer at his residence rather than on an OpenAI machine or network. Nelson Decl. Ex. 45 at 41:14–43:18; Choffnes Decl. Ex. A at ¶ 156.

151. Radford wrote a script that would have restricted the download to books listed for sale on Amazon.com, but he did not run that script. Nelson Decl. Ex. 738.

152. At the time of his download, Radford knew LibGen was "a Russian website" containing free books. Nelson Decl. Ex. 4 at 113:6–10; Nelson Decl. Ex. 45 at 113:17–22.

153. Radford knew he could have bought the books instead, and testified that downloading them from LibGen involved "no administrative burden." Nelson Decl. Ex. 45 at 113:7–22, 147:7–13, 147:23–148:13.

154. After the download finished, Radford transferred the books from his personal computer to an external hard drive, which he brought to OpenAI. Nelson Decl. Ex. 45 at 37:1–38:2.

155. Radford testified that he "brought that hard drive, not a copy of it, into OpenAI's office," and that afterward "[i]t just sat in my closet." *Id.* at 37:12–38:8.

156. One of Radford's colleagues "at OpenAI" then uploaded the books from that hard drive to an OpenAI company machine. *Id.* at 40:9–18, 41:2–3.

157. Later, when Dario Amodei asked on a draft OpenAI white paper, "Is this all of LibGen?", Tom Brown answered, "I believe that we added all books in LibGen that were also listed for sale on Amazon." Nelson Decl. Ex. 247.

158. Radford responded: "It's a moderately sized subsample of libgen - not the whole thing," and "Yeah it's books that string match ones on sale at amazon that had epub or mobi formats. The scraper wasn't fully finished when we merged the dataset in, either." *Id.*

159. OpenAI produced Radford's original LibGen downloads at OPCO_SDNY_1736877 to OPCO_SDNY_1736901. Shan Decl. Ex. A at ¶ 49. Plaintiffs' expert Shawn Shan identified 96 of Plaintiffs' Works in those files. Shan Decl. ¶ 31.

160. Testifying as OpenAI's corporate designee, Alex Paino confirmed that the acquisition producing the LibGen-1 dataset occurred in late 2018. Nelson Decl. Ex. 43(a) at 216:22–217:1. Paino confirmed that it was performed "by Alec Radford and Alec Radford alone." *Id.*

C. November 2018–February 2019: OpenAI Budgeted to Buy Books, Then Built Its First LibGen Dataset Instead

161. In November 2018, OpenAI revisited whether it should lawfully acquire books. Dario Amodei wrote that OpenAI should establish a [redacted] Nelson Decl. Ex. 333 at -640.

162. Brad Lightcap, OpenAI's former CFO and COO, responded: "On books, I'd like to scope it to understand what the $ impact would be if we moved forward with it in 2019… who can help me with this?" *Id.* at -639.

163. Dario Amodei responded: "David [Luan], Alec [Radford], and Daniela [Amodei] have state on it, I think." *Id.*

164. Lightcap replied: [redacted] Nelson Decl. Ex. 345.

165. Daniela Amodei responded: "So I talked to Alec and Dario and have some info for you: … Books costs $10/each – some may be more like $15," [redacted] *Id.*

166. Lightcap responded: [redacted] *Id.*

167. OpenAI did not move forward with buying books in 2019. Nelson Decl. Ex. 8 at 87:25–88:3.

168. On February 12, 2019, Mira Murati participated in discussions about how OpenAI could "get a wider range of books, either via freely available ones online or via a partnership." Nelson Decl. Ex. 293 at -989.

169. Murati suggested partnering with [redacted] *Id.* Murati noted that she had already [redacted] *Id.* Murati added that OpenAI could also [redacted] *Id.*

170. David Lansky, OpenAI's then-General Counsel, participated in the same discussion. Lansky suggested that OpenAI acquire "free book downloads [from] **libgen.io**," which had "an API and ways to do bulk downloads (e.g. https://github.com/x-mel/bigen)." *Id.* (emphasis added); Nelson Decl. Ex. 84.

171. Lansky testified that he learned of LibGen by "doing a Google search" while "generally searching for sources of books or other content," and that after he found the site and looked at it, he understood it to contain free books. Nelson Decl. Ex. 31 at 60:10–61:25. Lansky told Daniela Amodei that LibGen "existed and it is a potential source of content that we should explore," and "that it appeared to have content freely available and that there was a bulk download or API feature." *Id.* at 28:20–30:13.

172. Lansky testified that "I don't recall anybody objecting to my suggestion that they look at Library Genesis," and that he did not know whether OpenAI had any policy prohibiting the use of Library Genesis. *Id.* at 32:12–33:20.

173. Daniela Amodei circulated Lansky's "great ideas" "for how to get a wider range of books" to Radford, Murati, Lansky, Ilya Sutskever, and David Luan. Nelson Decl. Ex. 293 at -987 & -989.

174. Lansky reported he had also [redacted] and [] reached out to a best selling author who happens to be a big fan of OpenAI to see if he has any [redacted] connections." *Id.* at -988.

175. Lansky followed up that he had "some traction with [redacted]" and asked whether other employees were available to meet with them. *Id.* at -987.

176. Sutskever wrote that "getting [redacted] books is a super low hanging fruit and I think it's worth pushing pretty hard on getting them." *Id.* at -988.

177. OpenAI [redacted] On February 25, 2019, OpenAI employee Ben Chess copied the raw text out of the books Radford had downloaded and created a dataset originally called "libgen-reversible." Nelson Decl. Ex. 234 at -960; Shan Decl. Ex. A at ¶ 50. Chess added the dataset to "[redacted]" Nelson Decl. Ex. 234 at -960.

178. The dataset contained "117.5k books"—the exact number Radford had downloaded. *Id.* That dataset later became LibGen1. Shan Decl. Ex. A at ¶ 50.

179. OpenAI described the composition of that dataset in an internal document titled "Questions for Ben on data and data pipeline": "Libgen1: Alec collected one of them on a computer in his basement. He collected a fraction of the torrents, filtered for English, filtered for ISBN to make sure they are actual books." Nelson Decl. Ex. 300 at 747–48.

180. Shantanu Jain wrote that "libgen1 is an alec dataset, there are stories of him bringing a dusty hard drive from some closet to the office with it," and that "libgen1 is all english." Nelson Decl. Ex. 372 at -278. Radford testified that the download ran in his bedroom, not a basement, and that it was a direct download rather than a torrent. Nelson Decl. Ex. 45 at 42:22–43:21, 48:1–7, 49:5–16; Choffnes Decl. Ex. A at ¶ 150.

181. In early 2019, Radford proposed that OpenAI use the approximately 120,000 books he had downloaded from LibGen to train its models. Nelson Decl. Ex. 13 at 21:14–25.

D. February–July 2019: OpenAI Presented LibGen to Microsoft and Closed a $1 Billion Investment

182. In February 2019, OpenAI published the paper Language Models are Unsupervised Multitask Learners, which described the training of GPT-2. Nelson Decl. Ex. 77.

183. OpenAI's co-founder and CEO Sam Altman then secured a meeting with Microsoft's CEO Satya Nadella, and met with Nadella that very month, where Nadella made clear that Microsoft was interested in technology transfer and commercialization with OpenAI, and said the next step was to meet with Bill Gates. Nelson Decl. Ex. 2(a) at 28:7–17; Nelson Decl. Ex. 362 at -635.

184. On March 15, 2019, Lansky followed up again about downloading LibGen, asking: "did anyone ever dig in to the Library Genesis link I noted way at the bottom of this thread? At the bottom of this page you can download the most recent dump of their database (which is huge!)." Nelson Decl. Ex. 317.

185. On April 12, 2019, OpenAI's Board of Directors received a PowerPoint containing a Research Update from Bob McGrew, OpenAI's Vice President of Research, and Dario Amodei. Nelson Decl. Ex. 357 at -726; Nelson Decl. Ex. 2(a) at 255:21–256:12.

186. A slide titled "Language Results: GPT-2 Models" stated that OpenAI was "[t]raining 6B and 25B parameter model[s] on new datasets (Libgen, Book Corpus)." Nelson Decl. Ex. 357; Seeley Decl. Ex. A at ¶ 40.

187. In April 2019, OpenAI prepared a document for its meeting with Bill Gates and other Microsoft executives, including Umesh Madan, a technical advisor to Microsoft's CTO. Nelson Decl. Ex. 2(a) at 219:15-220:13; Nelson Decl. Ex. 188 at -958.

188. That document described the early results of OpenAI's training of the model eventually launched as GPT-3. Nelson Decl. Ex. 812. The document stated that the model had grown nearly tenfold in size from GPT-2 after OpenAI "added another ~11B words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context." *Id.* at 911–12.

189. A draft OpenAI–Microsoft presentation prepared for that meeting stated under "Data & Training" that OpenAI had trained "6B and 12B parameter model[s]" on "additional data sets, Book Corpus, LibGen, Wikipedia." Nelson Decl. Ex. 330 at -573; Nelson Decl. Ex. 2(a) at 226:24–227:5.

190. The document prepared for its meeting with Bill Gates linked LibGen's Wikipedia page. Nelson Decl. Ex. 812; Nelson Decl. Ex. 327 at -135. That page expressly noted that the United States District Court for the Southern District of New York had ordered the site shut down. Nelson Decl. Ex. 670.

191. Following the meeting with Gates, Microsoft agreed to invest $1 billion in OpenAI. Nelson Decl. Ex. 145 at -347; Nelson Decl. Ex. 203.

E. <u>September 2019–January 2020: Mann and Hallacy Torrented the Entire LibGen Collection</u>

192. After receiving Microsoft's $1 billion investment, OpenAI did not purchase or license books. Nelson Decl. Ex. 8 at 87:25–88:3; Lasinski Decl. Ex. A at ¶ 102.

193. Between September 2019 and January 2020, OpenAI employees Benjamin Mann and Chris Hallacy torrented approximately 35 terabytes of data from LibGen. Shan Decl. Ex. A at ¶¶ 17, 60; Choffnes Decl. Ex. A at ¶¶ 9, 154, 158; Nelson Decl. Ex. 33 at 145:2–11.

194. That volume equaled the entirety of LibGen's fiction and nonfiction book collections. Shan Decl. Ex. A at ¶¶ 17, 59, 60; Choffnes Decl. Ex. A at ¶¶ 78, 93, 94.

195. On September 18, 2019, Mann and Hallacy participated in a "Data-thon." Nelson Decl. Ex. 249. Mann created the Google Calendar invitation, added it to the "OpenAI Office Events" calendar, and sent it to at least Dario Amodei, Tom Brown, and Hallacy. *Id*.

196. The invitation directed participants to "[c]ollect" datasets and listed "LibGen" first, followed by "Guttenberg", "Open Library", "IRC", "Movie Scripts", and "anything you can think of!" *Id.*

197. Hallacy described a "Data-thon" as a "day-long effort . . . to acquire datasets." Nelson Decl. Ex. 25 at 105:11–14.

198. Two days later, on September 20, 2019, Hallacy received an email with the subject line "Welcome to Library Genesis: Miner's Hut." Nelson Decl. Ex. 250.

199. The email stated: "Welcome to Library Genesis: Miner's Hut forums[.] Please keep this email for your records. Your account information is as follows: Username: [redacted]" *Id.* Hallacy testified that "[redacted]" is "a handle I have used in other places." Nelson Decl. Ex. 25 at 107:1–13.

200. Later that day, Mann wrote in an OpenAI Slack channel that "[H]allacy and [I] have begun downloading libgen." Nelson Decl. Ex. 260.

201. Also on September 20, 2019, in another OpenAI Slack channel, Hallacy asked for help with "libgen parsing" and "the easiest way to add a persistent disk to a dev box"—that is, extra storage attached to a development computer. Nelson Decl. Ex. 308; Nelson Decl. 25 at 113:4–16.

202. When Chess offered a solution, Hallacy responded, "easy enough. I shall do that." Nelson Decl. Ex. 308.

203. Three days later, on September 23, 2019, Mann posted in an OpenAI Slack channel that "@Chris Hallacy produced some unfiltered extracted libgen samples!" Nelson Decl. Ex. 799; Nelson Decl. Ex. 25 at 115:15–25.

204. On September 25, 2019, Hallacy reported in an OpenAI Slack channel that "[w]e've also played around with libgen." Nelson Decl. Ex. 800. He had processed a single shard of the collection. He stated that "[a]t those rates we'd be hoping for 500GB of compressed text in all of libgen as an upperbound." *Id.* Hallacy added that the same storage location held "the torrent files needed to source the dataset" and "the first approx 200 shards of the dataset that I've downloaded." *Id.*

205. On October 3, 2019, Hallacy commented on a GitHub pull request from Mann for a "Spark-based domain classifier" that included code "for libgen" that would "help[] eliminate copyright info from start of document." Nelson Decl. Ex. 230.

206. That Spark-based classifier removed copyright information from documents downloaded from LibGen. Nelson Decl. Ex. 25 at 119:21–120:9.

207. Torrenting transfers data over a peer-to-peer network, in which participants both download files from and upload files to one another. Nelson Decl. Ex. 33 at 116:11–117:10.

208. On October 4, 2019, Mann posted a to-do list in an OpenAI Slack channel under the heading "Today: expand libgen." Nelson Decl. Ex. 802; Nelson Decl. Ex. 33 at 103:17–104:7. The list included "start headless transmission on ubuntu with blocklist and 0 upload," "create instance with 30 TB of storage," and "start downloading torrents." Nelson Decl. Ex. 802.

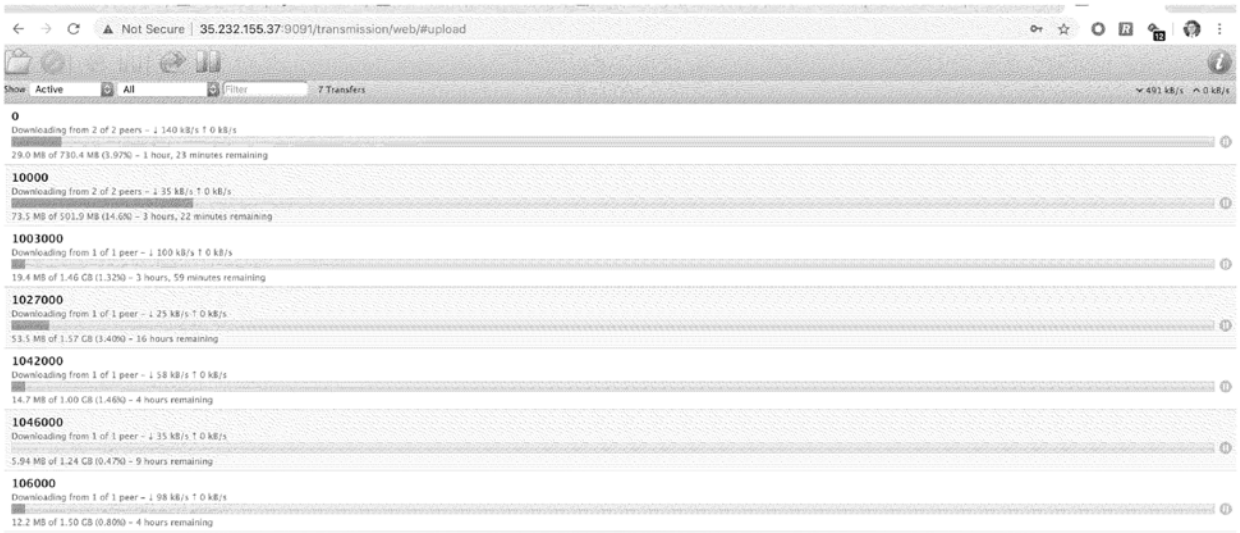
209. "Transmission" is the BitTorrent program OpenAI used to torrent. Choffnes Decl. Ex. A at ¶ 15. The "0 upload" setting was not the program's default; Mann changed the

configuration file to prevent OpenAI from sharing books back to other users while it downloaded. *Id.* at ¶¶ 127, 129.

210. On October 7, 2019, Mann reported to Dario Amodei, OpenAI researcher Sam McCandlish, and Brown that he had "start[ed] libgen torrent downloading." Nelson Decl. Ex. 267; Nelson Decl. Ex. 33 at 109:1–14, 114:14–19.

211. On October 8, 2019, in the same channel, Mann estimated that torrenting "all 3TB" would take approximately "a week." Nelson Decl. Ex. 778 at -525. He then revised that estimate to "2w given the average rate last night." *Id.*; Choffnes Decl. Ex. A at ¶ 73.

212. In the same OpenAI Slack channel, Mann posted a screenshot of the Transmission BitTorrent software managing the LibGen torrenting.



Nelson Decl. Ex. 268.

213. The screenshot shows seven of the LibGen torrents transferring. OpenAI ultimately queued more than 2,000 torrents. Choffnes Decl. Ex. A at ¶ 74. The torrent names visible in the screenshot—including "10000," "1003000," "1027000," "1042000," "1046000," and "106000"—follow LibGen's torrent naming convention. Under that convention, each torrent corresponds to an increment of 1,000 files. Shan Decl. Ex. A at ¶ 56; Nelson Decl. Ex. 268.

214. The screenshot also shows IP address 35.232.155.37, a Google-owned address, reflecting that OpenAI used Google Compute Engine to torrent. Choffnes Decl. Ex. A at ¶ 74.

215. The transfers ran through Google's IP address. Other users sharing the same files saw that address rather than an OpenAI address, which would have identified OpenAI as the party torrenting LibGen. *Id.*

216. On October 8, 2019, Mann posted in an OpenAI Slack channel that he was "trying to torrent libgen" but that "none of [his] torrents are downloading." Nelson Decl. Ex. 309 at -826. He asked how to expose a port to the public internet from inside his "rcall box," the internal computing environment he was using. *Id.*; Choffnes Decl. Ex. A at ¶ 75. Chess assisted Mann, and Mann got his torrents working. Nelson Decl. Ex. 309 at -827.

217. Later on October 8, 2019, Mann told Brown that he was "pulling the rest of libgen" and had "all 2176 torrents queued." Nelson Decl. Ex. 803. Each LibGen torrent file contained approximately 1,000 books; the 2,176 queued torrents therefore corresponded to more than 2.1 million books. Choffnes Decl. Ex. A at ¶¶ 16, 77.

218. On October 9, 2019, two days after Mann began torrenting, he submitted a pull request in OpenAI's GitHub repository titled "Libgen downloader and spark-based extractor (#4)." Nelson Decl. Ex. 329. A pull request is a proposed change to computer code. *Id.* One purpose of the code was to install Transmission, the torrenting software. *Id.*

219. The code shows that Hallacy stored the torrent files needed to download LibGen in his Google Storage folder. Nelson Decl. Ex. 25 at 143:2–144:3; Choffnes Decl. Ex. A at ¶ 133. Hallacy commented on that line of code, "Oh hey there's me." Nelson Decl. Ex. 329. The next day, Hallacy approved Mann's pull request. Nelson Decl. Ex. 804.

220. On October 11, 2019, Mann listed "[b]abysit libgen torrent" as one of his tasks. Nelson Decl. Ex. 805. He reported that he had torrented 544 gigabytes of books from LibGen. *Id.*; Nelson Decl. Ex. 33 at 139:16–140:16.

221. On October 11, 2019, Mann "read [A]lec's libgen code." Nelson Decl. Ex. 805. Mann testified that this was "the prior art for how to download and make use of LibGen that [he] was leaning on." Nelson Decl. Ex. 33 at 140:17–141:20. Mann posted a link titled "metadata-based libgen filter." Nelson Decl. Ex. 805.

222. On October 12, 2019, Brown asked Mann whether he planned to torrent all of LibGen or only LibGen fiction. Nelson Decl. Ex. 803. Mann answered that he was then torrenting only LibGen fiction but was investigating all of LibGen, which he reported was "10x bigger." *Id.* Brown replied that "[his] guess is that non-fiction will be even more useful" and offered additional OpenAI resources to aid the torrent. *Id.*

223. On October 16, 2019, Mann reported in an OpenAI Slack channel on the size of the LibGen nonfiction collection, which he was then evaluating. Nelson Decl. Ex. 373. It held 1.4 million PDF files and roughly 500,000 other documents, or approximately 2 million books. *Id.*; Choffnes Decl. Ex. A at ¶ 81. He posted that "libgen fiction is looking super high quality from the random sample (34 documents) i checked. didn't see a single bad entry." Nelson Decl. Ex. 373 at -583.

224. On October 24, 2019, Mann submitted a pull request to OpenAI's GitHub repository titled "Dedupe new libgen." Nelson Decl. Ex. 687. Mann stated that the dataset of books he had torrented was located in his Google Storage folder. *Id.*

225. On November 8, 2019, Dario Amodei committed a change to OpenAI's GitHub repository labeled "register libgen." Nelson Decl. Ex. 229.

226. On November 18, 2019, Mann submitted a pull request showing that he was downloading LibGen nonfiction to his persistent disk—the storage attached to the server performing the download. Nelson Decl. Ex. 688. That pull request marks the start of Mann's torrenting of LibGen nonfiction. Choffnes Decl. Ex. A at ¶ 91. That same day, in a document he shared with Radford, Mann posted a link to "new libgen," from which he took only mobi and epub file types. Nelson Decl. Ex. 301 at -624.

227. By December 4, 2019, Mann had torrented 7.3 terabytes of LibGen nonfiction. Nelson Decl. Ex. 806. Four OpenAI employees responded to Mann's message with the "astonished face" emoji. *Id.* Brown responded, "omg let's train on it!" and added five book emojis. *Id.* Mann answered, "haha ok i'll upload and extract." *Id.*

228. On December 11, 2019, Mann set as one of his objectives for the first quarter of 2020 filtering the new pull of LibGen and adding it to OpenAI's data mix. Nelson Decl. Ex. 257.

229. On December 12, 2019, Mann reported that he had "27GB of nonfiction libgen ready (4x more than epub/mobi only version[)]." Nelson Decl. Ex. 263.

230. On December 13, 2019, Mann reported that he had "accidentally included txt files" in the newest LibGen dataset. Nelson Decl. Ex. 256 at -338.

231. On December 17, 2019, Mann submitted a pull request to OpenAI's GitHub repository to "[u]pdate libgen fiction and . . . add bigger nonfiction. . . . [b]ecause PDFs handle whitespace strangely and when you convert to text things like 2 column layouts or page footers get all jumbled up." Nelson Decl. Ex. 231.

232. January 9, 2020 was a Learning Day at OpenAI—a recurring day on which OpenAI encouraged employees to learn about a topic relevant to OpenAI that interested them. Nelson Decl. Ex. 801; Nelson Decl. Ex. 25 at 189:1–5. Mann announced that his Learning Day plan was to

investigate PDF parsing techniques "so that we can better extract our 32TB of libgen nonfiction." Nelson Decl. Ex. 801. The post identified Hallacy as Mann's "Learning Day Buddy" for the project. *Id*.

233. On January 11, 2020, Hallacy received a GitHub notification with the subject line "Libgen nonfiction (#13)." Nelson Decl. Ex. 807. The body of the email read "Download complete! 32 TB" and included a link to a "New extraction with most of the data." *Id*. The LibGen nonfiction download was then complete. Choffnes Decl. Ex. A at ¶ 94.

234. OpenAI produced billing logs for the Google Compute Engine project it used to torrent. Choffnes Decl. Ex. A at ¶ 137. The logs showed monthly upload and download volumes during the download period. Dr. David Choffnes, Plaintiffs' expert, analyzed those logs to determine the volume of data transferred from the internet to the virtual machine that ran the LibGen torrents. *Id.*

235. Mann confirmed that he "torrented both [the] fiction and nonfiction" components of LibGen. Nelson Decl. Ex. 33 at 145:2–11.

236. The fiction collection comprised 2,182,642 book files, of which 865,087 were unique titles. Choffnes Decl. Ex. A at ¶ 78; Nelson Decl. Ex. 262.

237. The nonfiction collection totaled 32.38 tebibytes and more than 2 million books, which matched the LibGen nonfiction metadata as of January 2020. Nelson Decl. Ex. 801; Nelson Decl. Ex. 807; Choffnes Decl. Ex. A at ¶¶ 81, 93, 94; Shan Decl. Ex. A at ¶ 59.

238. Mann ultimately processed approximately 452,790 books from these torrents into the dataset known as LibGen2. Shan Decl. Ex. A at ¶ 62. He drew files from both the fiction and nonfiction portions of LibGen. Shan Decl. Ex. A at ¶ 62; Nelson Decl. Ex. 300 at -748.

239. An internal OpenAI document described that dataset as follows: "Libgen2: Two variants, fiction and non-fiction. Mostly used fiction. Difference here is all the epub and mobis were extracted either using caliber or pandoc. Otherwise trusting that everything else is good instead of doing lots of filtering." Nelson Decl. Ex. 300 at -748.

240. Jain later wrote that "libgen2 was made by ben mann because libgen1 was great but small," that "libgen2 seems mostly fiction," and that "libgen2 is 30% non-english." Nelson Decl. Ex. 372 at -278.

241. Across all of its LibGen downloads, OpenAI acquired millions of books. Shan Decl. Ex. A at ¶ 48.

242. As discussed in Section IV, OpenAI used only a fraction of those books to train its models. The LibGen1 and LibGen2 training datasets together comprised approximately 570,000 books. Millions of the books OpenAI downloaded from LibGen were never used to train any OpenAI model. *Id.* at ¶¶ 48, 50, 62.

243. Hallacy confirmed that OpenAI does not "use all of the datasets it downloads or acquires to train language models." Nelson Decl. Ex. 25 at 237:9–13.

F. 2019–June 2022: OpenAI Stored the Pirated Books in Locations Accessible to Its Employees and Put Them to Continued Use

1. Where OpenAI Stored the Books

244. OpenAI maintained LibGen in centralized storage for years. Testifying about the June 2022 deletion effort, OpenAI's Vice President of Research, Bob McGrew, explained that at the time OpenAI was [redacted] Nelson Decl. Ex. 35 at 112:24–113:10. McGrew testified: [redacted]

[redacted] *Id.* at 113:11–14. McGrew added: [redacted] *Id.* at 113:15–18.

245. Hallacy stored the approximately 86,984 books he downloaded from LibGen in 2019 in his own folder in OpenAI's Azure storage account—his "personal container"—labeled the "LibGen Test" file. Nelson Decl. Ex. 62(d) at 21:20–23:24, 30:8–16, 45:2–15; Shan Decl. Ex. A at ¶ 54.

246. Hallacy testified that these [redacted] where a dataset called LibGen also appears to have been stored. Nelson Decl. Ex. 25 at 228:1–230:4.

247. Hallacy used the LibGen Test file to create the unfiltered extracted LibGen samples he shared in OpenAI Slack channels. Nelson Decl. Ex. 799; Nelson Decl. Ex. 62(d) at 49:2–22. [redacted] Nelson Decl. Ex. 799; Nelson Decl. Ex. 62(d) at 49:2–55:12.

248. The LibGen Test file is organized into folders that follow the naming convention of LibGen torrents; each folder corresponds to an increment of 1,000 files. Shan Decl. Ex. A at ¶ 54; Nelson Decl. Ex. 828.

249. The files within those folders bear Md5 hashes that correspond to files LibGen distributed through the torrents matching the folder names. Shan Decl. Ex. A at ¶ 54. The metadata for the "1000" torrent of the LibGen fiction collection lists the Md5 hash "60b0dfd89390521464411d910d654ce7." Nelson Decl. Ex. 829 at 8.

250. A file bearing that name appears inside the "1000" folder of the LibGen Test file, and that file is a 571-page copy of Plaintiff George R. R. Martin's A Game of Thrones. Nelson Decl. Ex. 828 at 12; Nelson Decl. Ex. 830 at 10.

251. A second file named "78a6b22ec470da9ebf54c2ea98d22bb7" appears in the same LibGen fiction torrent and in the LibGen Test file, and is likewise a copy of A Game of Thrones. Nelson Decl. Ex. 829; Nelson Decl. Ex. 828 at 14; Nelson Decl. Ex. 831.

2. <u>Who Could Access the Books</u>

252. Jeff Wu testified that when he sought out LibGen-sourced materials at OpenAI he accessed them "from cloud storage," and that he "at least had access to some of the derived data, if not the original." Nelson Decl. Ex. 785 at 102:1–107:14.

253. Asked whether he was "aware of any access restrictions placed on access to LibGen data at any time during [his] tenure at OpenAI," Wu answered: [redacted] Nelson Decl. Ex. 785 at 107:15–24. Qiming Yuan testified that although [redacted] Nelson Decl. Ex. 67 at 145:1–145:25.

254. OpenAI's data access policy provides that [redacted] Nelson Decl. Ex. 361 at -377. OpenAI's dataset preparation code, [redacted] Shan Decl. Ex. A at ¶ 55.

255. Dr. Shan opines that "a personal container is not private storage: OpenAI's own dataset-preparation code [redacted] reads training data from another

employee's personal container, so hosting the books there left them broadly accessible." Shan Decl. Ex. F at ¶ 50.

256. Testifying as OpenAI's corporate designee, Alex Paino confirmed that "[a]nyone working to prepare a dataset would need to have access to the dataset in some form prior to its final form," that the LibGen datasets were originally "stored on BLOB storage," and that when he joined OpenAI in March 2019 the company "did not have a robust access-control system." Nelson Decl. Ex. 43(a) at 305:25–307:15.

3. <u>Copies, Migration, and Retention</u>

257. OpenAI made and retained multiple copies of the LibGen books in cloud storage. A file produced in this litigation, "libgen_overlap.ipynb," compares three LibGen collections in Google Storage. Choffnes Decl. Ex. A at ¶ 135. That file records 117,860 files in "original_libgen" at [redacted], 117,286 files in "libgen1" at [redacted], and 518,769 files in "libgen2" at [redacted]. *Id.*

258. A second file, "libgen_dedup_join.ipynb," references at least five additional LibGen storage locations. Choffnes Decl. Ex. A at ¶ 134.

259. OpenAI also had a dataset tracker listing "libgen-raw," "unfiltered_libgen," and "deduped_libgen," the latter two of which were stored in Mann's personal "rcall" directory: gs://ben-rcall/libgen/1021-clean/ and gs://ben-rcall/libgen/dedup-1026/. *See* Shan Decl. Ex. A at ¶ 72; Nelson Decl. Ex. 252.

260. In November 2020, Mann "migrated all [of] . . . raw libgen (fiction & nonfiction)" to OpenAI's Azure blob storage. Nelson Decl. Ex. 264; Choffnes Decl. Ex. A at ¶ 136.

261. As of June 2022, both the original and the derived versions of the data OpenAI downloaded from LibGen still existed on OpenAI's Azure blob storage. Nelson Decl. Ex. 338; Nelson Decl. Ex. 43(a) at 307:10–20; Nelson Decl. Ex. 62(b) at 66:12–67:8.

4. Continued Use of the LibGen Books

262. OpenAI used the LibGen data for purposes beyond pretraining. Those purposes included [redacted]. Nelson Decl. Ex. 266 at -255; Nelson Decl. Ex. 261 at -577; Nelson Decl. Ex. 253 at -924, -927; Nelson Decl. Ex. 43(a) at 237:2-247:23. OpenAI trained a [redacted] on LibGen [redacted] Nelson Decl. Ex. 43(a) at 243:2–246:21; Nelson Decl. Ex. 253.

263. In September 2019, Dario Amodei [redacted] like LibGen. Nelson Decl. Ex. 266 -at 255. Amodei directed Mann to "prioritize making it possible" to mix in common crawl with LibGen. *Id*.

264. In October 2019, OpenAI trained a new classifier designating the LibGen dataset as "good." Nelson Decl. Ex. 261 at -577. The LibGen dataset served as the positive training class against which OpenAI measured web-crawled data. *Id*. at -577.

265. OpenAI stopped using LibGen to train its publicly released GPT models in late 2021, but retained and continued to use LibGen thereafter. *See generally* Nelson Decl. Ex. 339; Nelson Decl. Ex. 340; Nelson Decl. Ex. 265 at -909.

266. In a September 2021 document titled "Language Data and Where to Find it," Jain theorized that, based on a "recent libgen metadata DB dump," OpenAI could "5x the amount of libgen data with PDFs." Nelson Decl. Ex. 306 at -467. Jain observed that "the best way to get more out of libgen is to solve PDF extraction." *Id*. Jain noted that "no work" had been done on the two LibGen datasets since their original creation. *Id*.

267. In February 2022, Long Ouyang wrote that "[redacted] because OpenAI has "a [redacted]" that it was not then training on. Nelson Decl. Ex. 240 at -897. Those books remained in OpenAI's storage. *See id*.

268. OpenAI was still using the LibGen data at the time it deleted that data. In June 2022, an OpenAI employee stated that OpenAI "actively us[ed]" the [redacted] dataset containing LibGen for "instruction following finetune experiments." Nelson Decl. Ex. 339 at -002. A June 2022 Slack message listed "libgen_1_dedup_clean" and "libgen_2_dedup_clean" among the datasets for instruction-following models. Nelson Decl. Ex. 265 at -909.

G. <u>April–May 2020: OpenAI Renamed LibGen "Books1" and "Books2"</u>

269. Mann, in partnership with other OpenAI employees, created the OpenAI dataset known as LibGen2, which came to be publicly identified as Books2. Nelson Decl. Ex. 33 at 66:4–11.

270. Internally, OpenAI called those two datasets "LibGen-1" and "LibGen-2." Nelson Decl. Ex. 45 at 39:10–40:8; Nelson Decl. Ex. 50(b) at 27:10–28:2; Nelson Decl. Ex. 43(a) at 232:10–232:15; Nelson Decl. Ex. 62(b) at 48:15–49:1.

271. Mann testified that "Books1 is a dataset that Alec Radford produced, and Books2 is a dataset that I produced, and both are based on LibGen," and that "internally at OpenAI, Books1 and Books2 were known as LibGen1 and LibGen2." Nelson Decl. Ex. 693 at 390, 393–94.

272. OpenAI's file and dataset names reflected the same origin. Nelson Decl. Ex. 62(b) at 49:9–24; Nelson Decl. Ex. 316. OpenAI named the version of Books1 it used to train GPT-3 "libgen1-dedup-clean." Nelson Decl. Ex. 62(b) at 49:9–24, 50:10–52:14, 144:3–145:4; Nelson Decl. Ex. 316. OpenAI named the version of Books1 it used to train GPT-3.5 "libgen1-dedup-clean-csam-filt." Nelson Decl. Ex. 62(b) at 49:9–24, 50:10–52:14, 144:3–145:4; Nelson Decl. Ex. 316.

273. OpenAI named the corresponding Books2 datasets "LibGen-2-dedup-clean" (used to train GPT-3), Nelson Decl. Ex. 57(b) at 36:13–15, and "libgen2-enonly-dedup-clean-csam-filt" and "libgen2-nonen-dedup" (used to train GPT-3.5), *see* Nelson Decl. Ex. 316. *See also* Shan Decl. Ex. A at ¶¶ 62–63.

274. In April 2020, on a draft of the paper that would announce GPT-3 publicly, Mann commented under the heading "Library Genesis": "given fuzziness on licensing for this dataset, do we want to just say 'a large corpus of books?'" Nelson Decl. Ex. 811; Nelson Decl. Ex. 33 at 255:4–11, 256:10–19. Radford left a "+1" reaction to Mann's question, a reaction OpenAI employees used to recognize a good question or a point for discussion. Nelson Decl. Ex. 811; Nelson Decl. Ex. 45 at 155:2–6. Melanie Subbiah commented on the same draft: "what are the rules around using this dataset in research and publications?" Nelson Decl. Ex. 811.

275. When Jack Clark, an OpenAI employee, asked Mann to add more specificity to the description of Books1 and Books2 in an early draft of the GPT-3 paper, Mann responded: "it's deliberately vague since it's libgen." Nelson Decl. Ex. 228 at -666.

276. On May 15, 2020, as OpenAI prepared the GPT-3 paper for publication, Dario Amodei raised "additional FUD/reputational risks to consider" in an OpenAI Slack channel. Nelson Decl. Ex. 371. Amodei asked: "Is it sketchy to call our corpuses 'Books1' and 'Books2' and not say what they are, particularly when in fact they are Libgen (which is a slightly sketchy source)." *Id.*

277. Radford replied that "Google in the Meena chatbot paper completely obfuscates where they got the data." Nelson Decl. Ex. 371. Radford added: "I think it is weird to have the specificity of Books1 and Books2 without talking more about them, it highlights vagueness by contrast." *Id*. Amodei asked, "what would you suggest here?" *Id*. Mann proposed, "we could merge

the token counts and just say books." *Id*. Radford responded, "yeah that's probably the minimum viable change." *Id*. at -558. Ashley Pilipiszyn asked, "can you call it WebBooks?" *Id*. Tom Brown, OpenAI's research engineering lead for GPT-3, replied, "WebBooks sounds reasonable!" *Id*. Sam McCandlish noted, "we call it 'Internet Books' in the scaling laws paper." Nelson Decl. Ex. 254 at -902.

278. Immediately before publication, Dario Amodei confirmed that OpenAI would "describe libgen as 'Books1' and 'Books2.'" Nelson Decl. Ex. 255 at -950.

279. OpenAI published the GPT-3 paper in May 2020 without disclosing LibGen. *See generally* Nelson Decl. Ex. 162; Nelson Decl. Ex. 45 at 210:14–211:24. The paper called LibGen1 and LibGen2 "Books1" and "Books2." Nelson Decl. Ex. 371; Nelson Decl. Ex. 45 at 211:11–24.

280. OpenAI did not obfuscate Wikipedia or Common Crawl as sources in the same paper. Nelson Decl. Ex. 33 at 310:20–311:7.

281. OpenAI witnesses did not provide a non-privileged explanation for the renaming. Nelson Decl. Ex. 693 at 394–97; Nelson Decl. Ex. 4 at 217:25–218:24; Nelson Decl. Ex. 50(b) at 28:3–12; Nelson Decl. Ex. 785 at 82:14–18; Nelson Decl. Ex. 62(b) at 49:2–8.

282. Mann testified that he "was part of discussions in which that was discussed." Nelson Decl. Ex. 693 at 394–97. Mann declined to answer why OpenAI chose "Books1" and "Books2" instead of "LibGen1" and "LibGen2" on the ground that answering would reveal "the substance of conversations between myself and attorneys." *Id*. Dario Amodei testified that he did not "remember" why "OpenAI changed the name of the internal dataset from LibGen-1 and LibGen-2 to Books1 and Books2." Nelson Decl. Ex. 4 at 217:25–218:24. Nicholas Ryder, an OpenAI model and data scientist for GPT, testified: "I wasn't involved in those decisions." Nelson Decl. Ex. 50(b) at 28:3–12. Jeff Wu, an OpenAI research engineer, testified that he did not

"remember." Nelson Decl. Ex. 785 at 82:14–18. OpenAI's corporate designee testified: "I don't have any awareness of that." Nelson Decl. Ex. 62(b) at 49:2–8.

H. April 2021–2024: OpenAI Repeatedly Priced Lawful Alternatives and Never Replaced the LibGen Books

283. In April 2021, OpenAI created a Slack channel called "#large-book-dataset," to which Sam Altman was alerted, to "coordinate on data needs for text" and to "figure out a way to get access to a very large book corpus" for "future GPT training runs." Nelson Decl. Ex. 694; Nelson Decl. Ex. 695; Lasinski Decl. Ex. B at ¶ 115. Greg Brockman stated that he wanted "all the books in the world" or the best approximation OpenAI could get, and that "as much diversity of data as we can get" was preferred. Nelson Decl. Ex. 694; Lasinski Decl. Ex. B at ¶ 115.

284. On May 28, 2021, Jain wrote to Brockman that the "#large-book-dataset is moving kind of slowly. [I]f we want more books, I think we might be able to get more out of libgen than we have currently, particularly if we get a pdf to text." Nelson Decl. Ex. 370.

285. In 2021, Sutskever wrote that [redacted]" Nelson Decl. Ex. 358. John Schulman responded that [redacted] *Id*.

286. Schulman added: "I guess we could buy them to cover ourselves legally, and also download a copy on libgen, we don't need to buy the ebooks either, [we] can buy physical copies." *Id*.

287. In 2021, Radford wrote that [redacted] Nelson Decl. Ex. 694. Radford estimated that [redacted] *Id.* at -095.

288. Radford described himself as "bullish" on the "direct" routes because they "won't involve protracted discussions / negotiations / licensing" and OpenAI could "get started today."

> AR
> Alec Radford
> so bought for $20 each off amazon seems pretty doable? 4/6/2021, 11:01 PM UTC
>
> I think I'm bullish on these "direct" routes because they won't involve protracted discussions / negotiations / licensing and we could "get started today" but understand I'm pretty naive on the other org decision factors involved. 4/6/2021, 11:05 PM UTC

Id.

289. Brockman asked whether Radford "would automate the purchasing." *Id.* Radford responded that having contractors do it would be fine but that "it seems like it's possible to script this up on something like amazon." *Id.*

290. Radford noted that most books on Amazon carried Digital Rights Management software. *Id*. Radford stated: "I guess an even sketchier rout [sic] if DRM is blocking is to purchase a copy of every book but still use the DRM cracked version already available through something like libgen." *Id*. Radford stated that the "purchase for yourself route [would] work." *Id*.

291. In January 2022, Brockman wrote that OpenAI should license books for artificial intelligence training, stating that "if we can pay to play right now we probably should." Nelson Decl. Ex. 311.

292. [redacted] Lasinski Decl. Ex. A. at ¶ 102. [redacted] Nelson Decl. Ex. 698. [redacted] *Id*.

293. In 2022, OpenAI evaluated the cost of buying books to replace those it had downloaded from LibGen. Nelson Decl. Ex. 321. In private messages with Brockman, Jain stated [redacted] to replace [the approximately 670,000 books Jain estimated made up the two] libgen[] [datasets]" and [redacted] to match gopher," Google's model, which Jain estimated had 6.8 million books in its training data. *Id*. Jain described this as the "book buying approach." *Id*.

294. Later in 2022, Brockman wrote to Sutskever that "books is like the best long-context data . . . cleanest, least risky." Nelson Decl. Ex. 347 at -550. Brockman wrote that "we can also purchase a lot of books, but they are expensive per token so we haven't prioritized." *Id*. at -551. The same exchange reflects that negotiating book agreements with publishers required "effort" and could be "expensive." *Id*. at -551–52.

[redacted] In 2024, OpenAI investigated purchasing books, considering agreements to bulk buy and scan books. Lasinski Decl. Ex. A at ¶ 105. In April 2024, OpenAI signed [redacted] with [redacted] Nelson Decl. Ex. 699. Under that agreement, [redacted]

296. Varun Shetty, OpenAI's Vice President of Media Partnerships and corporate representative, testified that he was not aware of OpenAI ever purchasing books at scale and scanning them for purposes of training data: "No, I'm not aware of that." Nelson Decl. Ex. 55(a) at 233:25–234:2.

297. Qiming Yuan testified that OpenAI should have been willing to pay "10X of the cost of training a model" for a "high-quality, diverse e-books dataset." Nelson Decl. Ex. 67 at 158:11–160:6. Yuan testified that the "token acquisition cost should be higher than the compute cost" because "you can reuse the data that you acquired." *Id.*

298. In 2021 and early 2022, OpenAI contemplated gaining access to [redacted]. Nelson Decl. Ex. 695; Lasinski Decl. Ex. B at ¶ 115 & n.439. In that discussion, Lightcap observed that [redacted] "likes to push the limits of copyright." Nelson Decl. Ex. 695.

299. OpenAI explored obtaining books from [redacted] but did not obtain any. Nelson Decl. Ex. 8 at 176:8–177:19. Brockman confirmed that OpenAI "[redacted] *Id*. at 124:14–18.

300. Brundage testified that a proposal to obtain books through a partnership was treated as low priority because OpenAI already had LibGen. Asked about a message stating, "We have talked to them, but the books that would be most interesting to us would require a partnership. We're currently getting lots of books through LibGen, so it seems low priority," Brundage agreed that this "does seem to be what Ben Mann is saying." Nelson Decl. Ex. 9 at 290:14–24.

301. OpenAI tested whether public-domain books could substitute for the LibGen books and determined that they could not. Nelson Decl. Ex. 289. Qiming Yuan scraped "[redacted] *Id*. at -367. Yuan reported in October 2021 that "[m]ixing in public domain books seems to have negative impact on sampling eval. By mixing in public domain books at same ratio as libgen 1 and 2, we see slight degradation on sampling eval performance." *Id*.

302. Yuan determined that "[i]t doesn't seem like public domain books can replace lib[gen] without losing performance," reporting "about ~3% drop on [redacted] where libgen has biggest impact." *Id*. Yuan stated: "Overall I feel it's probably ok to include public domain books as a separate dataset but I would view it as lower quality than libgen due to much older books and high OCR errors." *Id.*; Nelson Decl. Ex. 67 at 149:25–155:24.

303. OpenAI's experiments showed that substituting public domain books for the LibGen books required more compute and produced a model that, while functional, performed worse. Nelson Decl. Ex. 289; Nelson Decl. Ex. 67 at 149:25–155:24; Nelson Decl. Ex. 280.

304. OpenAI sought access to the books that [redacted] held but [redacted] refused. Nelson Decl. Ex. 311 at -706. OpenAI employee Jason Kwon stated that OpenAI "[redacted]" *Id*. Kwon stated that [redacted] was "[redacted]." *Id*.

I. <u>December 2021–Summer 2022: OpenAI Deleted LibGen</u>

305. In late 2021, Schulman wrote to Brockman, Sutskever, and other OpenAI employees regarding a new paper from Anthropic reporting their success on model training that "presumably [Anthropic] did a much more thorough scrape of libgen. I wonder if we should just do that too, and override the lawyers." Nelson Decl. Ex. 323 at -922.

306. On December 3, 2021, at 4:14 p.m. UTC, Sutskever asked in an OpenAI Slack channel: "what's the state with the lawyers re libgen?" *Id*. at -923.

307. Schulman replied at 4:17 p.m.: "we're not planning to use [LibGen]." *Id*. Brockman asked at 4:19 p.m.: "who made that call btw?" *Id*.

308. Wojciech Zaremba, an OpenAI co-founder, responded at 4:20 p.m. that OpenAI was "about to use [redacted] which contains not copyright books" and was "working on getting all the books from other legal sources." *Id*.

309. Schulman responded at 4:25 p.m. that he was "not optimistic about getting the books from legal sources, and [redacted]." *Id*.

310. OpenAI meeting notes from December 2021 state: "can't use libgen due to legal reasons." Nelson Decl. Ex. 320 at -476.

311. On January 2, 2022, Brockman asked other OpenAI employees: "I know we're ditching libgen, but are we keeping the public domain books?" Nelson Decl. Ex. 840.

312. On June 15, 2022, in an OpenAI Slack channel, Paino asked: "general q: how concerned are we about mentions of libgen? (they're all over google docs/slack/github . . .)". Nelson Decl. Ex. 222. That evening, McGrew wrote to Paino and Nicholas Ryder: "Given how much OpenAI is in the news, now is the right time to excise Libgen from our systems and storage. What would be involved in that?" Nelson Decl. Ex. 338; Nelson Decl. Ex. 35 at 98:22–100:04. McGrew identified three categories of material to remove: "Direct storage of libgen," "Derived datasets," and "GitHub storage." Nelson Decl. Ex. 338; Nelson Decl. Ex. 785 at 101:23–102:21; Nelson Decl. Ex. 43(b) at 142:6–14. Paino then added Lilian Weng, Jeff Wu, and Jakub Pachocki to the channel to locate the data. Nelson Decl. Ex. 338 at -003.

313. McGrew wrote in the same thread that "removing libgen entirely would stop us from being able to repro GPT-3 or GPT-3.5 again (but would be very valuable for legal reasons)." Nelson Decl. Ex. 338 at -002. McGrew asked whether there were "any upcoming experiments where we would like to repro GPT-3." *Id.*

314. McGrew testified that Project Clear "looks like it was a project to remove the dataset Libgen from OpenAI's systems and storage." Nelson Decl. Ex. 35 at 98:20–99:22. Paino testified that Project Clear was "the name given to an effort at this time to remove some data from our systems." Nelson Decl. Ex. 43(b) at 141:24–142:5. Paino testified that "data from LibGen is included in the set of things that here Bob is talking about deleting and would like deleted." *Id.*

315. OpenAI employees originally named the Slack channel "excise-libgen." Nelson Decl. Ex. 338; Nelson Decl. Ex. 35 at 115:11–20. OpenAI's then-General Counsel, Jason Kwon, renamed it "project-clear." Nelson Decl. Ex. 338; Nelson Decl. Ex. 35 at 115:11–20. McGrew testified that, "outside of conversation with counsel," he had "no independent idea why Jason would have renamed it that way." Nelson Decl. Ex. 35 at 115:14–16.

316. The #project-clear channel remained active through at least June 29, 2022, when Paino, McGrew, Kwon, Weng, and Wu continued to discuss removing LibGen references and dataset paths. Nelson Decl. Ex. 222.

317. McGrew directed that OpenAI "remove all these components by July 1st" and asked, "what do we need to do to make that happen?" Nelson Decl. Ex. 338; Nelson Decl. Ex. 35 at 115:21–116:6. McGrew testified that he referred to "any direct copies of the Libgen dataset that might exist, any derived datasets that exist and if there were any . . . copies of that dataset that was in the GitHub cloud storage." Nelson Decl. Ex. 35 at 116:2–6.

318. Lilian Weng, OpenAI's then-Vice President of Research for Safety, responded, "I can start removing data right now. Shall I?" Paino answered, "yeah I think that'd be great! and good point re: [redacted], forgot we still had libgen in there." Weng asked, "This includes both libgen-1 and libgen-2 right?", and Paino answered, "yep." Nelson Decl. Ex. 338 at -001–02.

319. Jeff Wu wrote in the same channel: "the books project used a small subset of libgen as a source. we don't really need the datasets anymore. should i just delete the original source dataset or also all the books human data that uses some subset of libgen? the latter might be a little harder to track down completely." Nelson Decl. Ex. 222 at -001; Nelson Decl. Ex. 785 at 109:1–110:25.

320. Paino compiled a list of the datasets containing LibGen books that OpenAI had determined to delete. Nelson Decl. Ex. 222 at -003. The list spanned approximately 18 pages and contained more than 498 entries. Nelson Decl. Ex. 340.

321. When OpenAI considered removing the LibGen data from its systems, Paino wrote that "the [redacted] and [redacted] mixes are the main ones we still care about which uses libgen," and that "[i]f we want to fully excise libgen, then we'll just have to accept that we can't repro [redacted]." Nelson Decl. Ex. 338 at -001.

322. Paino then cautioned: "Actually maybe we should hold off on deleting the final filtered version of libgen1/2 used for [redacted] until we decide we're OK with not being able to repro those results." *Id.* at -002.

323. Paino confirmed that "we deleted at least some of the Lib-Gen derived data." Nelson Decl. Ex. 43(a) at 272:12–13. McGrew testified that he did not know whether the deletion "was complete by July 1st," but that "eventually it was . . . removed." Nelson Decl. Ex. 35 at 119:24–120:1.

324. The LibGen datasets are the only training datasets OpenAI has ever deleted. Nelson Decl. Ex. 62(a) at 59:7–60:8.

325. Because of these deletions, OpenAI has not produced the original set of files that Mann downloaded from LibGen in 2019, any logs from Mann's torrenting software, or the versions of the LibGen datasets it used to train GPT-3.5. Choffnes Decl. Ex. B at ¶¶ 8–10, 83, 88.

326. OpenAI's corporate representative Michael Trinh testified that the only copies of Books1 and Books2 that OpenAI now has are those used to train GPT-3, and that OpenAI "haven't been able to find files that match . . . the second set—the UUIDS that we believe were used for GPT-3.5." Nelson Decl. Ex. 62(b) at 147:3–15. OpenAI has never located the version of the dataset used to train GPT-3.5—"libgen1-dedup-clean-csam-filt"—following the summer 2022 deletion. Shan Decl. Ex. A at ¶ 51.

327. OpenAI deleted the LibGen data. The deletion prevented OpenAI from reproducing its earlier model training runs. Nelson Decl. Ex. 380 ("One thing we'll get burned by is the ability to not reproduce earlier results. . . . It means that deleting data from GCS has to be a big deal. If earlier models rely on that data and it disappears out from under them, we'll have no ability to recreate it."); Nelson Decl. Ex. 43(b) at 152:6–153:24. OpenAI witnesses testified that reproducibility mattered to their research. Nelson Decl. Ex. 67 at 23:18–24:5; Nelson Decl. Ex. 43(b) at 152:6–153:24.

328. OpenAI has not identified any non-privileged reason for deleting the LibGen materials, and has not asserted any solely practical or technical reason for the deletion. *See* ECF No. 1276 at 3–4; Nelson Decl. Ex. 62(a) at 56:13–58:8, 59:23–60:8; Nelson Decl. Ex. 62(b) at 122:13–124:23.

J. <u>Throughout, OpenAI Knew LibGen Was "Sketchy [as F---]"</u>

329. OpenAI and Microsoft employees at every level understood throughout this period that LibGen was a piracy site and that piracy is wrong and illegal. *See infra* §§ II.J.1-2.

1. <u>OpenAI's Contemporaneous Knowledge</u>

330. In February 2016, Jay Smith informed Brockman that "there are legal questions around [the] use" of "sites like Library Genesis (LibGen)," which offer "free access to a large range of publications." Nelson Decl. Ex. 220 at -022. Smith added that "neither the Wikimedia Foundation nor the Wikipedia community endorses them." *Id.*

331. In 2019, Miles Brundage, OpenAI's Head of Policy Research, acknowledged that his legal concerns about OpenAI's activities included "copyright concerns." Nelson Decl. Ex. 9 at 273:21-274:5.

332. In 2019, Wu stated that he was "unclear whether the methods of originally obtaining [LibGen] were legal and also unclear about the legality of training on those datasets." Nelson Decl. Ex. 785 at 91:10–25.

333. On February 19, 2019, Chess emailed Daniela Amodei, Radford, Dario Amodei, and OpenAI Vice President of Engineering David Luan. Chess noted that "we discussed several book datasets at the team meeting a couple weeks ago" and flagged "legal issues that we needed to consider" regarding those datasets. Nelson Decl. Ex. 337. Those datasets included the LibGen books Radford downloaded and kept "in his closet." Nelson Decl. Ex. 337; Nelson Decl. Ex. 13 at 61:19–62:12.

334. On March 28, 2019, Dario Amodei pasted into an OpenAI Slack thread the LibGen Wikipedia entry's description of the site's database, quoting that it "contains more than 2.7 million books and 58 million science magazine files." Nelson Decl. Ex. 384; Nelson Decl. Ex. 809.

335. In April 2019, OpenAI leadership, including Altman, Dario Amodei, Brockman, Murati, and Sutskever, prepared a demonstration and written materials for Bill Gates and other Microsoft executives. *See* Nelson Decl. Ex. 810; Nelson Decl. Ex. 287.

336. On April 18, 2019, Altman shared with Gates the April 2019 document described above, which stated under the heading "Data" that the model had grown nearly tenfold in size from GPT-2 after OpenAI "added another ~11B words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context." *See* Nelson Decl. Ex. 810; Nelson Decl. Ex. 813; Nelson Decl. Ex. 812 at 911–12; Nelson Decl. Ex. 2(a) at 217:23–218:22; *see supra* § II.D. This document was also shared with other executives at Microsoft, including Chief Technology Officer Kevin Scott. Nelson Decl. Ex. 702; Nelson Decl. Ex. 703 at -984–85.

337. As of April 2019, the Wikipedia page that OpenAI linked in its document described LibGen as allowing "free access to content that is otherwise paywalled or not digitized elsewhere." Nelson Decl. Ex. 670; Nelson Decl. Ex. 384 ; Nelson Decl. Ex. 385. The page stated that "[i]n 2015, the website became involved in a legal case when Elsevier accused it of providing pirate access to articles and books." *Id*. The page stated that LibGen "is blocked by a number of ISPs in the United Kingdom." *Id*. The page stated that "[i]n late October 2015, the District Court for the Southern District of New York ordered LibGen to shut down and to suspend use of the domain name (libgen.org), but the site is accessible through alternate domains." *Id*.

338. Dario Amodei testified that "[a]nyone who had read the Wikipedia article would have read that" LibGen has been "accused of providing pirate access to articles and books." Nelson Decl. Ex. 4 at 158:22–159:14.

339. On May 24, 2019, Schulman shared a link to LibGen in an OpenAI Slack channel, accompanied by the following text: "Library Genesis is a scientific community targeting collection of books on natural science disciplines and engineering." Nelson Decl. Ex. 814. The link shows that Schulman was searching for textbooks about linear algebra. *Id*. Schulman had downloaded

textbooks from LibGen in college, and knew that LibGen contained books with "varying legal rights." Nelson Decl. Ex. 52 at 99:20–102:12, 106:2–20.

340. Schulman recognized the legal exposure that LibGen created, remarking that it was a "good thing we're not actually selling books from LibGen." Nelson Decl. Ex. 323.

341. On July 19, 2019, McCandlish wrote in an OpenAI Slack channel: "We're not sure if we're going to release the Foresight LM Scaling paper publicly, but if we do we were thinking about removing all mentions of LibGen, since it's a bit of a sketchy data source." Nelson Decl. Ex. 297. Dario Amodei, OpenAI's then-Research Director, responded that "as a training set [LibGen is] a bit sketchier." *Id.* McCandlish explained: "I was just worried about optics - i.e. 'openai uses copyrighted data from sketchy russian website' showing up on HN would be unfortunate." *Id.* at -016.

342. Chess questioned the legality of LibGen on a draft of OpenAI's paper Scaling Laws for Neural Language Models. The draft stated that OpenAI "also test[ed] on similarly-prepared samples of Library Genesis." Nelson Decl. Ex. 684 at -447. Chess commented: "Is LibGen legit?" *Id.*

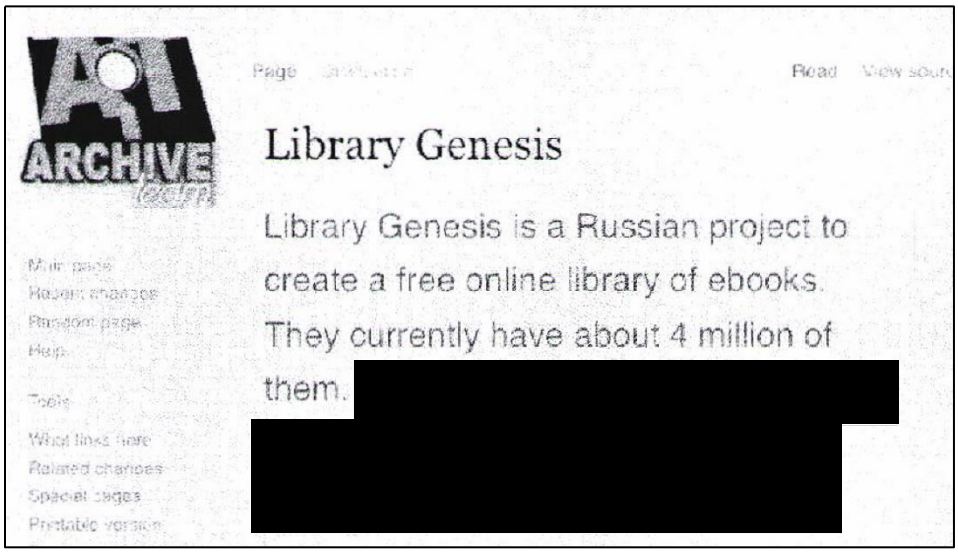
343. In the published version of the scaling laws paper, OpenAI described the LibGen data as "Internet Books" rather than LibGen. Nelson Decl. Ex. 233; Nelson Decl. Ex. 684.

344. In internal notes dated August 19, 2019, an OpenAI employee wrote: "We trained GPT-3 on pirated stuff! No sharing that!" Nelson Decl. Ex. 302 at -705.

345. On September 14, 2019, Mann told Chess that he had "loaded our libgen into [redacted]" an OpenAI storage database, and posted a link to a webpage describing LibGen. Nelson Decl. Ex. 816; Nelson Decl. Ex. 13 at 114:21–115:4. The webpage stated: "Library Genesis is a Russian project to create a free online library of ebooks. They currently have about 4 million

of them. The project is in blatant violation of copyright, and their domain name is blocked in some countries." Nelson Decl. Ex. 841.

346. A screenshot of that same webpage, found in the personal files of OpenAI employee Tom Brown, bears the same statement that the project "is in blatant violation of copyright."



Nelson Decl. Ex. 817.

347. Mann testified that before he torrented LibGen he had heard others describe Library Genesis as a pirated website, and that "I knew that LibGen was shut down." Nelson Decl. Ex. 33 at 261:4–19, 273:2. Mann answered "that's right" when asked "What you did to download LibGen, you were doing it with the knowing consent of others at OpenAI; correct?" *Id.* at 277:5–8.

348. Mann testified that "LibGen was sketchy [as f---]." *Id.* at 77:18–78:15.

349. On October 4, 2019, Hallacy drafted a document titled "Data Plan for Algorithms." Nelson Decl. Ex. 380. Under the heading "other considerations," Hallacy wrote: "Once AGI is hit we're gonna want to commercialize. [Microsoft] already wants a commercial product. Our datasets have to come from commercially licensed sources for that model." *Id.* at -049. Tracked changes show that an earlier version of the draft read "legit sources" in place of "commercially licensed

sources." *Id.* Hallacy wrote that "[i]t's also not clear if we can commercialize products obtained via scraping." *Id.* at -055.

350. In December 2019, Hallacy commented on a GitHub pull request from Mann and explained that a processing step was useful because "[f]or libgen this helps to eliminate copyright info from start of document." Nelson Decl. Ex. 230 at -864. Hallacy participated in developing that processing step. Nelson Decl. Ex. 25 at 119:21–120:9; *see supra* Section II.E.

351. In February 2020, Dario Amodei told Brockman and others that he "wish[ed] we had an operation called ''forget training data'" because such an operation "relates to privacy and copyright issues we're likely to encounter in commercialization eventually." Nelson Decl. Ex. 818, at -125.

352. Notes from an OpenAI meeting on April 17, 2020, attended by Mann and others, state that "Libgen is sketchy [as f---]."

- No code, vague on data
  - Libgen is sketchy AF

Nelson Decl. Ex. 819 at -894.

353. OpenAI directed that a forthcoming paper contain "no code, vague on data." Nelson Decl. Ex. 819 at -894. That direction appears in the same April 17, 2020 meeting notes recording that "Libgen is sketchy [as f---]." *Id.*

354. OpenAI researchers, including Tom Brown, described LibGen as "sketchy AF." Nelson Decl. Ex. 819 at -894. Mann and other OpenAI employees confirmed that "AF" stands for "as f---." Nelson Decl. Ex. 33 at 77:18–20.

355. In July 2020, another OpenAI employee wrote: "it does seem nice to try to buy the books, and maybe it's worth thinking about how much of a pain that would be[.] 10,000 books is

probably too many though … I think we just won't say where the books are from." Nelson Decl. Ex. 246 at -242.

356. In July 2020, OpenAI employee Ryan Lowe assessed the risk of continuing to use LibGen for the book-summarization project. Lowe wrote that he thought "there's a >80% chance that we have some exchange of the form: 'where did you get the books data?'" and "'we can't say'[.]" Nelson Decl. Ex. 325 at -315. Lowe estimated "a further ~40% chance that that leads to a moderate-sized Twitter kerfuffle that negatively affects the external perception of our work." *Id.* Lowe added: "if we're fully okay with these potential outcomes, then I'm comfortable continuing using Libgen for the project." *Id.*

357. During that conversation concerning "OpenAI's policy towards training on copyrighted data," OpenAI employee Daniel Ziegler stated that "I guess the fair use question is not totally settled in copyright law." Nelson Decl. Ex. 325 at -315–16. Ziegler asked: "What happened to the idea of just buying the books we end up training on?" *Id.* at -315.

358. In August 2020, Lowe wrote internally about his "FUD [fear, uncertainty, and doubt] about our use of Libgen for this project," noting that "we can't mention the source of the data." Nelson Decl. Ex. 223 at -682. A year later, in August 2021, a colleague asked Lowe: "how come we can't talk about our training data?" Lowe answered: "because we used libgen." Nelson Decl. Ex. 226 at -165.

359. On October 25, 2020, Andrew Mayne, an OpenAI employee and published author, observed that LibGen contained his own copyrighted works. Nelson Decl. Ex. 307.

360. In April 2021, a reporter asked OpenAI to confirm whether GPT-3 was trained on Books1 and Books2, whether the public had access to any of their contents, and how OpenAI assembled them. Nelson Decl. Ex. 276. An OpenAI employee told others that they "shouldn't be

saying anything" on the question whether Books1 and Books2 were public, and that OpenAI does not disclose anything beyond what already appears in the GPT-3 paper. *Id*.

361. In May 2021, Brundage sent a link to an article to Radford, Wu, and other OpenAI employees in a Slack channel finding "evidence that . . . BooksCorpus likely violates copyright restrictions for many books[.]" Nelson Decl. Ex. 788. Radford responded that "Only GPT-1 was trained on BooksCorpus . . . we were also aware of all these issues back then. I like to joke that half of GPT-1's training data was vampire romance novels." *Id.* Wu added that "we're using libgen . . . (which probably has the same problems)." *Id*. Another OpenAI employee reacted to Wu's message with a "facepalm" emoji. *Id.*

362. On November 30, 2023, OpenAI employees discussed in the private OpenAI Slack channel "#ai-news" whether competitors were training on LibGen. Nelson Decl. Ex. 324 at -317. One employee wrote that when x.ai launched, "a common speculation was that they'll just train on all of libgen." *Id*. Another employee wrote that he "strongly suspect[ed]" a second competitor was doing the same. *Id*. OpenAI employee Nat McAleese then wrote: "Train on libgen, raise money, get new data, delete the old models, be clean forever more. Temporal regulatory arbitrage." Nelson Decl. Ex. 324. Three colleagues reacted to McAleese's message with a ":not-legal-but-very-cool:" emoji. *Id.*

363. Another OpenAI employee in the same channel asked, "What's libgen?" Nelson Decl. Ex. 324 at -317. Three colleagues reacted to that question with the ":not-legal-but-very-cool:" emoji. *Id.* A colleague answered by linking the Library Genesis entry on Wikipedia and adding, "see also Sci-Hub, Z-Library and Annas Archive." *Id.*

364. OpenAI removed LibGen references from its code. Nelson Decl. Ex. 785 at 94:21–95:6. Wu testified that OpenAI undertook "an effort to remove references to certain datasets,"

which he "believed to be sourced from Library Genesis," including "an effort to remove references in code to certain . . . dataset paths." *Id*.

365. OpenAI removed references to LibGen from papers before publishing them. Nelson Decl. Ex. 297 at -015–16; Nelson Decl. Ex. 233. Eric Chung received an OpenAI paper that identified LibGen and forwarded it to others within Microsoft. When asked whether it surprised him that OpenAI "scrubbed down all references to Library Genesis" when it "released this paper publicly," Chung responded that he didn't "have any feelings about it." Nelson Decl. Ex. 14 at 60:20–61:15.

366. Ouyang described the LibGen data as "pirated books." Nelson Decl. Ex. 348.

367. Tasha McCauley, a former member of OpenAI's Board of Directors, testified that it would have mattered to her as a board member "that LibGen had been shut down by United States courts and nevertheless OpenAI was using material from LibGen to train its models." Asked whether it would have mattered had she known while serving on the board, McCauley answered: "That would've mattered to me, I believe." Nelson Decl. Ex. 834 at 132:10–19.

2. <u>Defendants' Admissions That Piracy Is Wrong and Illegal</u>

368. OpenAI employee Tao Xu testified that Library Genesis "might contain pirated books." Nelson Decl. Ex. 815, at 24:18–22. Xu agreed that, in general, "you shouldn't" go to a website known to host pirated books and obtain them for free rather than buying them. *Id.* at 83:16–84:5.

369. Shetty testified that it would not be fair for companies like OpenAI to access pirated websites like Library Genesis to obtain training material without paying for it, explaining: "I don't think you should be training on stolen material." Nelson Decl. Ex. 55(b) at 49:6–15.

370. Altman admitted: "Q: You agree that piracy is wrong and illegal? A: I do." Nelson Decl. Ex. 2(a) at 254:20–22; *id*. at 167:15–17 ("Q: [D]o you understand that piracy is illegal? A: I do."). Altman testified that it "doesn't feel good" to train on pirated data. *Id*. at 163:2–5.

371. Altman testified that if a person is "intentionally going to a website where you know the stuff is pirated and download that, that is something that we intend to not to do and I don't think we should [d]o." Nelson Decl. Ex. 2(a) at 168:8–16. Altman answered that if he learned that behavior was occurring at OpenAI, he would "[r]emove data from the training set and figure out how we need to change our policies or our procedures for that not to happen." *Id*. at 168:17–23.

372. Altman testified that it would not be "appropriate" to obtain a pirated copy of a book even if the person intended to write a review of that book, because the initial act of piracy would be "illegal" on its own. *Id*. at 167:8–14.

373. Dario Amodei, OpenAI's former Research Director, agreed that it would be wrong for a person to "go to Library Genesis and download a copyrighted book for free without paying for it in order to read that book" absent an applicable exception. Nelson Decl. Ex. 4 at 161:2–17. When asked whether he would obtain a copy of George R. R. Martin's *A Game of Thrones* or John Grisham's *The Pelican Brief* from a pirated website or buy it at a bookstore. Amodei answered: "I would buy it at a bookstore." *Id.* at 169:1-12.

374. Daniela Amodei, OpenAI's former Vice President of Safety and Policy, testified that it is not appropriate "to go to an illegal pirate website and take data from there for free." Nelson Decl. Ex. 3 at 43:11–15.

375. Chess testified that he "would believe piracy is "illegal" and that "downloading copyrighted material . . . for free" is illegal. Nelson Decl. Ex. 13 at 20:2–25, 23:17–23. Chess also testified that he had downloaded copyrighted content from The Pirate Bay to obtain "material that

I couldn't otherwise buy or obtain in some way," and agreed that he was aware at the time "that it was illegal to do it." Nelson Decl. Ex. 13 at 91:13–92:14.

376. Ilya Sutskever, OpenAI's co-founder and former Chief Scientist, testified that he could not identify "a meaning of the word 'pirated' that would not include a dataset like the one that we're discussing." Nelson Decl. Ex. 58 at 151:7–13.

377. Greg Brockman, OpenAI's co-founder and President, testified that he "generally do[es] not believe" that stealing a book from a bookstore is a legal source. Nelson Decl. Ex. 8 112:19–23.

378. OpenAI employee Tarun Gogineni wrote that piracy is "mostly just stealing." Nelson Decl. Ex. 821. In the same exchange, Gogineni wrote that piracy is "not conserving resources by not buying [the media in question], you're mostly just stealing[,] that being said i pirate shit all the time." *Id.*

379. When Gogineni recommended a television program and a colleague asked "where to watch it sir," he responded: "idk pirate it [to be honest]." Nelson Decl. Ex. 822.

380. During his tenure at OpenAI, Gogineni stated that it was "impossible" to find a copy of a book he wanted either "legally or illegally." Nelson Decl. Ex. 823.

381. Mann, the OpenAI employee who torrented the LibGen collections, testified that he downloaded books, movies, and software from the piracy site The Pirate Bay. Mann testified that he "do[es] sometimes visit The Pirate Bay" and that he had visited the site as recently as "the last month" before his deposition. Nelson Decl. Ex. 33 at 232:5–233:15. Mann agreed that "downloading copyrighted material without permission [could] be a crime." *Id.* at 233:17–21.

382. Satya Nadella, Microsoft's Chief Executive Officer, admitted: "Q: It's illegal to download pirated material, right? A: Absolutely. That's correct." Nelson Decl. Ex. 40 at 175:13–

19. Nadella described piracy as a "serious, industry-wide problem[]." *Id.* at 172:24–173:4. Nadella stated that Microsoft would "ideally strive to shut down all piracy" and confirmed that Microsoft should "not use pirated sites to obtain data." Nelson Decl. Ex. 40 at 195:4–6, 300:7–11.

383. Nadella agreed that "the same issues with respect to piracy occur with respect to books" as with software, and that "anybody who visits one of these sites and downloads pirated copyrighted material is at risk of legal liability." Nelson Decl. Ex. 40 at 173:20–23, 174:20–175:4.

384. Nadella testified that he "wish[ed] that OpenAI had not gone to a pirated site" and that he "would regret using a pirated site." Nelson Decl. Ex. 40 at 302:8–303:1. Nadella further testified that he would "separate the two"—the acquisition of books from a pirated site and the later use of books to train a model—and that OpenAI "could have scanned the copy," as "Google did with Google Books," rather than "go[ing] to a pirated site." *Id.*

385. Bill Gates, Microsoft's co-founder, testified that piracy in the copyright context "means to use a copyrighted material without paying for it." Nelson Decl. Ex. 21 at 19:7–13. Gates testified that "in almost every country, there is a copyright law that says that if you don't pay, you're breaking the law." Nelson Decl. Ex. 21 at 19:14–19. Gates agreed that piracy is illegal in the United States. Nelson Decl. Ex. 21 at 9:21–20:2.

386. Umesh Madan, technical advisor to Microsoft's Chief Technology Officer, agreed that downloading copyrighted works from a piracy site to avoid paying for them would be illegal. Nelson Decl. Ex. 32 at 56:9–15. Madan testified that he buys his books and CDs, that it "never occurred to me" to download pirated content, and that he "would not steal them." Nelson Decl. Ex. 32 at 57:8–58:2.

387. Microsoft's Corporate Vice President Saurabh Tiwary, who led some of Microsoft's early AI initiatives, agreed that it would not be appropriate to source training material from pirated websites. Nelson Decl. Ex. 61 at 35:16–36:4.

388. Mustafa Suleyman, Microsoft's Chief Executive Officer of AI, testified that Microsoft has "policies prohibiting the use of pirated data sources like LibGen." Suleyman confirmed that "ultimately, the decision was Microsoft will not use LibGen to train models," and that the reason was that doing so "was a violation of our policies" "[a]gainst using pirated content." Nelson Decl. Ex. 824 at 110:25–112:16.

389. Jordan Usdan testified that when Microsoft began training its own language models, he asked "what are our policies?" and "our policy was clear to not train on pirated content." Nelson Decl. Ex. 63(a) at 219:2–220:6.

390. Misha Bilenko, a Microsoft corporate vice president, testified that he "would not" personally obtain a book, movie, or game from a pirate website because "I believe it's unethical" and "piracy is not a good thing." Nelson Decl. Ex. 837 at 62:24–63:18.

391. Brent Hecht testified that he "would not download" content from Library Genesis "for a number of reasons one of which is . . . I don't want to use pirated content." Nelson Decl. Ex. 26 at 373:14–22.

[redacted]

III. OpenAI Also Downloaded Books from Books3, Targeted Scrapes, and Web Crawls

A. Books3 and The Pile

393. The Pile is a dataset that includes a component called "Books3," which Shawn Presser and Leo Gao assembled from the shadow library Bibliotik and announced in October 2020.

Nelson Decl. Ex. 307; Nelson Decl. Ex. 730, at -621 § 3.3; Nelson Decl. Ex. 672, at 3–4 § 2.3; Nelson Decl. Ex. 784, at 1.

394. Presser stated that Books3 derived from "all of Bibliotik" and contained 196,640 books. Nelson Decl. Ex. 307.

395. The copyright holders of those books did not consent to their inclusion on Bibliotik. Nelson Decl. Ex. 20 at 24:22–25:20.

396. Gao, a member of EleutherAI, compiled The Pile and was the first-listed author on the paper announcing it. Nelson Decl. Ex. 672, at 1; Nelson Decl. Ex. 29 at 314:20–315:10. Gao referred to the books in Books3 as "yarrharr data." Nelson Decl. Ex. 786; Nelson Decl. Ex. 20 at 103:5–20.

397. Gao testified that naming the dataset "Books3" was "certainly" a reference to the "shady 'Books1' corpus" that OpenAI had identified as training data for GPT-3. Nelson Decl. Ex. 730; Nelson Decl. Ex. 20 at 83:18–23, 133:15–22.

398. Gao testified that his "decision to explore books as a source of data was inspired by what I saw in the GPT-3 paper," and that he had been "frustrated with OpenAI's choice to no longer publish as much as they used to about their research." Nelson Decl. Ex. 20 at 72:24–74:14, 116:6–117:5.

399. In announcing Books3, Presser wrote that in "OpenAI's papers on GPT-2 and 3, you'll notice reference to datasets named 'books1' and 'books2' . . . . OpenAI will not release information about books2. A crucial mystery." Nelson Decl. Ex. 664 (https://x.com/theshawwn/status/1320282153595396096).

400. By October 23, 2020, the books3.tar.gz file was statically hosted for public download at the-eye.eu. Nelson Decl. Ex. 784 at 1. Gao responded on EleutherAI's public repository that he would "work on adding all the newly-hosted datasets in the next few days." *Id*.

401. On October 23, 2020, OpenAI employee Christian Gibson posted in an OpenAI Slack channel that "a (let's call it fringe-legal) data hoarding community that I belong to has just taken up the mantel of storing and distributing Eleuther AI's The Pile(TM)," and posted a graphic describing the contents of The Pile, which included Bibliotik. Nelson Decl. Ex. 244; *see* Shan Decl. Ex. A at ¶ 79.

402. On October 25, 2020, Sutskever posted a link to Presser's Books3 announcement in an OpenAI Slack channel, describing it as "possibly good training data." Nelson Decl. Ex. 307.

403. Shantanu Jain confirmed that The Pile was created by individuals "broadly affiliated with a collective known as Eleuther.AI," that Books3 "is a dataset of books derived from a copy of the contents of the Bibliotik private tracker," and that the paper's first author is the same Leo Gao whom OpenAI thanked for contributing to data for OpenAI's models. Nelson Decl. Ex. 29 at 313:23–315:24); Nelson Decl. Ex. 672 at 1, 3 § 2.3.

404. Plaintiffs' expert analysis shows that OpenAI downloaded at least one, and likely more than one, full copy of Books3 and of The Pile as a whole. Shan Decl. Ex. A at ¶¶ 80–108; Shan Decl. Ex. F at ¶ 28.

[redacted]

[redacted]

407. OpenAI employees also learned of The Pile and its contents through the Discord channel that members of EleutherAI used, in which at least Jeff Wu was a participant. Nelson Decl. Ex. 237; Nelson Decl. Ex. 20 at 161:25–165:24.

408. Gao did not work at OpenAI when the paper announcing The Pile was published. OpenAI hired him in late 2021. Nelson Decl. Ex. 668; Nelson Decl. Ex. 20 at 12:10–13:5. At OpenAI, Gao suggested that OpenAI train its models on The Pile datasets, including Books3. Nelson Decl. Ex. 20 at 145:1–146:12; Nelson Decl. Ex. 225 at -729–30. Gao understood that using LibGen "might get me sued." Nelson Decl. Ex. 673 at 86; Nelson Decl. Ex. 20 at 15:17–16:10.

409. Gao stated that "book data was really important for language understanding tasks," and OpenAI employees discussed how "Books3 compare[d] to our libgen datasets." Nelson Decl. Ex. 341; Nelson Decl. Ex. 285 at -159.

410. At OpenAI, employees use personal storage accounts—also known as personal containers—to "store[] personal copies of datasets . . . that they were working with in order to prepare them for their individual research efforts." Nelson Decl. Ex. 62(c) at 18:9–19:8.

411. On February 1, 2021, Vedant Misra, a member of OpenAI's Reasoning and Algorithm team, downloaded The Pile for research and submitted a pull request to OpenAI's GitHub repository adding The Pile to OpenAI's data repository, "[redacted]" and fixing EleutherAI's path name. Nelson Decl. Ex. 277; Nelson Decl. Ex. 37 at 36:16–37:25. At his deposition, Misra read into the record that "Books3 is a dataset of books derived from a copy of

the contents of the Bibliotik private tracker made available by Shawn Presser." Nelson Decl. Ex. 37 at 24:16–25:9; Nelson Decl. Ex. 238 at -621 § 3.3.

412. Another OpenAI employee [redacted] Nelson Decl. Ex. 665; Nelson Decl. Ex. 37 at 41:7–8; Nelson Decl. Ex. 43(a) at 285:19–286:3. Misra downloaded the data into OpenAI's Azure blob storage container. Seeley Decl. Ex. A at ¶ 51; Shan Decl. Ex. A at ¶ 106.

413. In the February 2021 notes of OpenAI's "Reasoning" team, Misra linked his pull request and stated that he "added Eleuther's 'The Pile' dataset to [redacted] (825 GB! Including PDFs!)." Nelson Decl. Ex. 331 at -701; Shan Decl. Ex. A at ¶ 81; Nelson Decl. Ex. 43(a) at 285:24–286:10.

414. The Pile totaled 825.18 GB, and Misra testified that "[s]omeone who downloaded the full Pile would get 825 gigabytes." Nelson Decl. Ex. 37 at 46:15–47:21; Nelson Decl. Ex. 238 at -621 tbl. 1.

415. Misra used The Pile to train models at OpenAI. Nelson Decl. Ex. 331 at -697–99); Nelson Decl. Ex. 375, at -872. Brockman monitored Misra's progress, asking "btw what's our status on testing out The Pile?" Nelson Decl. Ex. 259 at -234. Misra replied that he "[s]hould have a comparison this week-next" and that the "[i]nitial signs are positive." *Id.*

416. OpenAI included full copies of other datasets sourced from The Pile in the training data for [redacted] *See* Shan Decl. Ex. A at ¶ 103. Two Pile datasets called [redacted] were included in the [redacted] training data, *see* Nelson Decl. Ex. 310 at -710, and OpenAI's code from [redacted] references "the_pile" as a "[redacted]" for training, *see* Nelson Decl. Ex. 312.

417. In September 2023, OpenAI employee Jeff Wu downloaded The Pile to his personal container, submitting a pull request titled "download the pile (PR #47861)." Nelson Decl. Ex. 313; Nelson Decl. Ex. 314. Wu's personal container contains a file called "thepile.py" that references an internal dataset stored on Azure at [redacted] Shan Decl. Ex. A ¶ 108. EleutherAI had removed Books3 from its own hosted version of The Pile in August 2023, but other versions of The Pile available at that time still contained Books3. *Id*. at ¶ 109. OpenAI has not recovered or produced Wu's copy of The Pile. *Id*. at ¶¶ 108–09.

B. <u>OpenAI Deleted Its Copy of The Pile in October 2024</u>

418. Misra downloaded The Pile into his personal storage account and left OpenAI in 2021. Nelson Decl. Ex. 37 at 12:22–23, 19:1–9; Nelson Decl. Ex. 62(c) at 42:10–43:8.

419. Class Plaintiffs identified the Books3 component of The Pile as a recreation of the Bibliotik collection in their First Consolidated Amended Complaint, filed in February 2024. First Consolidated Am. Compl. ¶ 121.

420. In October 2024, OpenAI deleted the personal storage accounts of former employees, including Misra's account and the account of OpenAI co-founder and former Chief Scientist Ilya Sutskever. Nelson Decl. Ex. 62(c) at 41:6–22, 42:10–43:9, 43:24-44:4, 58:25-59:3.

421. OpenAI did not search for or produce any data from Misra's personal storage account before deleting it. Nelson Decl. Ex. 62(c) at 57:5–17, 58:3–8.

422. At the time of the deletion, OpenAI [redacted]. Nelson Decl. Ex. 62(c) at 36:21–38:25, 42:10–43:9, 45:11–46:18, 54:24–55:10.

423. When OpenAI deleted Misra's personal storage account, it lost the entirety of the data in that account and did not store that data anywhere else. Nelson Decl. Ex. 62(c) at 43:2–9, 46:4–6.

424. OpenAI retained no index of the files in the account and is not aware of any restoration of the deleted data. *Id.* at 43:16–44:5.

C. OpenAI Also Acquired Books Through Targeted Scrapes and Third-Party Web Crawls

425. OpenAI also acquired copies of books from sources other than LibGen and The Pile, including targeted "scrapes" of particular websites and third-party "crawls" of the internet as a whole. Shan Decl. Ex. A ¶¶ 111–12, 115–17.

426. A targeted scrape is a download of content available on a specific website. *See* Shan Decl. Ex. A ¶ 42. OpenAI performed a number of targeted scrapes, including scrapes of links posted by users of Reddit.com, YouTube videos, and podcast recordings, with the video and audio content processed using text-to-speech software. The resulting datasets include "[redacted]" Nelson Decl. Ex. 77 at 3; Nelson Decl. Ex. 351 at -872, -894.

427. Web crawlers are automated programs that visit websites and download their content. Shan Decl. Ex. A ¶ 111. AI companies use crawlers to collect data for internal storage and later use, including as training data for LLMs; to augment AI-backed assistants; and to facilitate

AI-backed search engines or search results. Many AI companies, including OpenAI, have developed their own crawlers. *Id.*

428. OpenAI uses the crawler "GPTBot" to collect and store web data for potential later use in training, the crawler "ChatGPT-User" to augment the AI assistants it offers through ChatGPT, and the crawler "SearchBot" to augment search results displayed to users in ChatGPT. Nelson Decl. Ex. 72 at 2.

429. Crawler names such as "GPTBot," "ChatGPT-User," and "SearchBot" are also known as "user agents," and they are typically how a bot appears to the websites whose content it crawls. Nelson Decl. Ex. 72 at 2. Companies that run traditional search engines, meaning those that do not exclusively provide results with the assistance of an LLM, such as Google or Bing, also operate crawlers to index the web incident to providing search results. *See id.*

430. AI companies including OpenAI also source crawled data from third parties, most prominently Common Crawl. Founded in 2007, Common Crawl describes itself as a "nonprofit 501(c)(3) organization that crawls the web and freely provides its archives and datasets to the public." Nelson Decl. Ex. 80.

431. Common Crawl states that it has downloaded more than 300 billion pages over 19 years and adds between 3 and 5 billion new pages each month. Nelson Decl. Ex. 79; Nelson Decl. Ex. 336 at -294.

432. Common Crawl operates the crawler known as "CCBot." *See* Nelson Decl. Ex. 72 at 4 (https://arxiv.org/pdf/2411.15091).

433. Common Crawl's "Terms of Use" state that Common Crawl "strongly recommends that you obtain the advice of legal counsel before making any use, including commercial use, of the Service and/or the Crawled Content" and that "BY USING THE CRAWLED CONTENT,

YOU AGREE TO RESPECT THE COPYRIGHTS AND OTHER APPLICABLE RIGHTS OF THIRD PARTIES IN AND TO THE MATERIAL CONTAINED THEREIN." Nelson Decl. Ex. 81.

434. Website operators may request that their pages not be crawled by specific crawlers by using "robots.txt." *See* Nelson Decl. Ex. 72 at 1–3 (https://arxiv.org/pdf/2411.15091). A website owner seeking exclusion from the Common Crawl would add the "CCBot" user agent to a "disallow" list. *Id.* A copyright owner whose work is hosted without consent on a website the owner does not control, such as an illegal shadow library, cannot modify that site's "disallow" list to prevent crawling of the pages containing the work. *See id.*

435. Plaintiffs' expert Dr. Shawn Shan is a co-author of the study cited in the preceding paragraphs, which he describes as "one of the first comprehensive technical analyses of the 'crawling' infrastructure used by AI companies to scrape training data from the internet." Shan Decl. Ex. A at ¶ 4.

436. Dr. Shan and his colleagues analyzed more than 50,000 popular websites over time to measure the adoption of robots.txt protocols directed to AI crawlers, and they performed passive and active technical measurements to determine which commercial AI crawlers respect blocking directives and which ignore them. Shan Decl. Ex. A at ¶ 4; Nelson Decl. Ex. 53 at 60:12–61:4.

437. Dr. Shan opines that opinions regarding robots.txt "do[] not apply to Plaintiffs' works which OpenAI pirated from LibGen, Books3, or obtained via scraping/crawling," because Plaintiffs, who are not the owners or administrators of the sites where their works are posted without consent, "do not have the ability to edit the site's robots.txt (or any other setting) to prevent crawling / scraping by OpenAI or Microsoft." Shan Decl. Ex. D at ¶ 51.

438. For the Accused Models, OpenAI obtained its third-party crawls primarily from Common Crawl. Shan Decl. Ex. A at ¶¶ 112–17. OpenAI has copied content from Common Crawl's crawl of the internet since at least 2019 and has used that content extensively, including as LLM training data. *Id.* at ¶ 117; Nelson Decl. Ex. 57(a) at 283:13–17 ("We have used the Common Crawl dataset to train various models.").

439. The training datasets derived from the raw data OpenAI downloaded from Common Crawl are generally denoted by a "[redacted]" prefix or by "[redacted]" elsewhere in the dataset name. *See* Nelson Decl. Ex. 351 (identifying, e.g., "[redacted] and [redacted] as sets used to train GPT-3; "[redacted]" as a set used to train GPT-3.5; "[redacted]" as a set used to train GPT-3.5 Turbo; "[redacted]" as a set used to train GPT-4; "[redacted]" as a set used to train GPT-4 Turbo; and "[redacted]" as a set used to train GPT-4o).

440. Common Crawl includes pirate sites such as b-ok.org, b-ok.cc, b-ok2.org, e-libra.ru, ubooki.ru, freenovel24.com, silo.pub, epdf.pub, and pdfsecret.com among the sites it crawls and downloads copies from. Shan Decl. Ex. A at ¶ 115.

441. Testifying on behalf of OpenAI, OpenAI employee Alex Paino acknowledged that, as part of its Common Crawl downloads, OpenAI had downloaded at least 56,000 documents from b-ok. *See* Nelson Decl. Ex. 43(a) at 211:24–212:21; Nelson Decl. Ex. 318 at -695; Nelson Decl. Ex. 690 at -019.

442. OpenAI employee Vinnie Monaco compiled the list of 56,000 b-ok documents and portions of the text downloaded and testified that the files came from OpenAI's training datasets. *See* Nelson Decl. Ex. 662 at 199:14–200:18.

443. Monaco testified that he compiled the list because he sought to identify instances of copyrighted book content in the training data. *See* Nelson Decl. Ex. 662 at 175:20–181:7, 199:14–200:18; Nelson Decl. Ex. 689.

444. Slack conversations between Paino, Monaco, and other OpenAI employees show that they understood b-ok to be a mirror of the shadow library Z-library, which the FBI shut down and whose founders were arrested. *See* Nelson Decl. Ex. 318 at -695; Nelson Decl. Ex. 798 at -461; Nelson Decl. Ex. 733.

445. OpenAI has also contributed financially to Common Crawl. In September 2023, OpenAI signed a contract to donate [redacted] to Common Crawl in three annual payments of [redacted]. *See* Nelson Decl. Ex. 343; Nelson Decl. Ex. 692; Nelson Decl. Ex. 343 (OpenAI employees noting the value OpenAI had gotten from Common Crawl in the past and "[w]ondering if we can pay more to let them scale up their crawl," with the employee responsible stating "they're going to do whatever we want").

446. The donation document stated that the grant was to further purposes that included "[i]ncreas[e] the cadence of [Common Crawl's] crawling to improve recency," "[i]ncreas[e] the crawl size - both breadth and depth," "[b]etter deduplication and spam avoidance," and "[i]mprove crawl strategy to seek out the highest-quality pages in each run." Nelson Decl. Ex. 692 at -032; Nelson Decl. Ex. 57(a) at 280:19–281:16.

447. Each of those purposes would increase the volume, recency, and quality of the data OpenAI downloaded from Common Crawl. Nelson Decl. Ex. 57(a) at 281:17–284:2.

IV. <u>OpenAI Trained Its LLMs on Datasets Containing Books</u>

A. <u>OpenAI Knew Books Were High-Quality Training Data</u>

448. In June 2018, OpenAI published a paper titled "Improving Language Understanding by Generative Pre-Training," which introduced OpenAI's first large language model, later called GPT-1. Nelson Decl. Ex. 78.

449. The paper states that OpenAI trained the model on a collection of books called "BooksCorpus," containing "over 7,000 unique" unpublished books from "a variety of genres including Adventure, Fantasy, and Romance." *Id.* at 4.

450. The paper further states: "Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information." *Id.* at 4. The paper contrasts BooksCorpus with an alternative dataset that was "shuffled at a sentence level – destroying long-range structure." *Id.* at 5.

451. Ben Chess, an OpenAI engineering manager, testified that "the thesis of OpenAI and its models was that by giving the more high-quality text that you feed into a model, the better it's going to perform, the better the results." Nelson Decl. Ex. 13 at 41:11–15.

452. Asked whether he understood that one type of high-quality data was books, Chess testified, "Seems reasonable to me," and that "it seems reasonable to think that would be one possible source" of high-quality text. Nelson Decl. Ex. 13 at 41:16–42:1.

453. Testifying as OpenAI's corporate designee, Nicholas Ryder stated that "[b]ooks are an example of a dataset that . . . contain very long, uninterrupted, correlated sequences of words," and that "in the early part of OpenAI, people were very preoccupied about this feature of books." Nelson Decl. Ex. 50(a) at 89:20–24.

454. In November 2018, OpenAI wrote a document that set forth "Top level goals." Nelson Decl. Ex. 284.

455. The first was to "build a large new text corpus for pretraining," which Dario Amodei, OpenAI's then-Research Director, commented was a "very high priority (as well as likely to succeed)." *Id.* at -140.

456. In late 2018, "it was a high priority for OpenAI to scale up all aspects of large language model training." Nelson Decl. Ex. 4 at 46:3–5.

457. In April and May 2019, OpenAI prepared a draft paper on the [redacted]-parameter model it then called GPT-3. Nelson Decl. Ex. 283.

458. Section 2.2 of that draft, titled "Expanding the Dataset," stated: "In order to scale up the model size by 8x without severely overfitting, we expanded the dataset beyond the 6 billion word dataset used in [Radford et al., 2019]," and that OpenAI "added another 11 billion words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context," bringing the combined datasets to "a grand total of [redacted] words." *Id.* at -066.

459. The draft further stated that OpenAI "removed from LibGen all books that were not also listed for sale on Amazon.com in order to increase the dataset quality," and that, although OpenAI found the webtext2 data had "a broader variety of content and topics than the LibGen data," "the LibGen data was still useful," so OpenAI "undersampled the LibGen data while still using the full dataset[.]" *Id.*

2.2 Expanding the Dataset

In order to scale up the model size by 8x without severely overfitting, we expanded the dataset beyond the 6 billion word dataset used in <<Radford et al, 2019>>. First, we simply collected more of the same type of data used in <<Radford et al, 2019>> (Reddit outlinks with at least 3 karma), resulting in a 15 billion word corpus having roughly the same distribution as the original dataset (we call this corpus "webtext2"). In addition, we added another 11 billion words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context. We removed from LibGen all books that were not also listed for sale on Amazon.com in order to increase the dataset quality. Finally, we added roughly 1B words from Wikipedia. Mixed together these datasets give a grand total of 27B words. Qualitatively we found that the webtext2 data had a broader variety of content and topics than the LibGen data, leading to better transfer on average, although the LibGen data was still useful. Therefore, in training, we undersampled the LibGen data while still using the full dataset -- specifically, the LibGen data was sampled only [redacted] of the time (despite being [redacted] of the data). This allowed for a good compromise between emphasizing the most useful data (webtext2) and having a lot of data.

[redacted]

461. In a 2020 document addressing model performance per unit of computing power, OpenAI stated that "[a] low-risk way to improve perFLOP for large language models is to improve the training data." Nelson Decl. Ex. 305.

462. In the paper announcing GPT-3, published in 2020, OpenAI stated that the CommonCrawl dataset was large enough "to train our largest models" but was of lower quality than "more curated datasets." Nelson Decl. Ex. 162 at -660. OpenAI stated that it filtered CommonCrawl based on how similar the data appeared to "a range of high-quality reference corpora," including Books1 and Books2. *Id.*

463. OpenAI further stated that it had "added several curated high-quality datasets, including an expanded version of the WebText dataset, collected by scraping links over a longer

period of time . . . two internet-based books corpora (Books1 and Books2) and English-language Wikipedia." *Id.*

466. In a document discussing which datasets to include in a 2020 model training run, Greg Brockman, OpenAI's co-founder and President, asked: "should we try to obtain more books data? [I] am definitely interested in the idea of an AI that has read all books . . . there is a lot of wisdom in them." Nelson Decl. Ex. 272. Ben Mann, an OpenAI member of technical staff, responded that OpenAI had [redacted] of books" but noted potential difficulties extracting the text. *Id.*

467. Ilya Sutskever, OpenAI's co-founder and Chief Scientist, responded: "If we can easily double the number of our training books on top of libgen it could be quite useful, as data quality will be major source of wins that don't cost any compute." Nelson Decl. Ex. 355; Nelson Decl. Ex. 251.

468. Brockman then emailed Sutskever asking, "[r]emind me do you buy into the idea of getting non-libgen books for world knowledge, though perhaps not perplexity?" Nelson Decl. Ex. 272. Sutskever replied: "yes if we could double the number of our books as a result. Do you

think that's possible? Improved data quality will be a major source of wins that won't require extra compute." *Id.*

469. An OpenAI document titled "GPT Wishlist" requested "more books (e.g. [redacted]),"[redacted] stating that books "are usually higher quality than websites" and that additional books "will help with [redacted]." Nelson Decl. Ex. 269.

470. A separate brainstorming document prepared by Nicholas Ryder listed "[t]rain on all the books in the world" under the heading "[n]ew data sources." Nelson Decl. Ex. 270.

471. Ben Mann, one of the OpenAI employees who originally downloaded LibGen, agreed that "books are invaluable for long-range context modeling research and coherent storytelling." Nelson Decl. Ex. 33 at 264:3–18; Nelson Decl. Ex. 301 at -613 (October 2020 notes of a one-on-one between Mann and Alec Radford stating "anything with long coherent text is high quality," with the first bullet underneath stating "watch libgen").

472. In discussing OpenAI's training data filtering pipeline, OpenAI researchers stated that the filter was "rejecting fairly high-quality books." Nelson Decl. Ex. 296 at -981.

473. In a Slack thread discussing the PDF portion of the Common Crawl web crawl, OpenAI researcher Alex Paino asked: "[redacted]" Nelson Decl. Ex. 239.

474. OpenAI researcher Tao Xu responded that "[redacted], likely accounting for similar or more tokens." Nelson Decl. Ex. 239; Nelson Decl. Ex. 8 at 213:25–214:20.

475. In December 2021, after Google published a paper describing a model called "Gopher," an OpenAI researcher analyzing the paper wrote: "the amount of books data stands out to me. they claim everything is publicly accessible, so we should find what books they're using

and use them too! they mention they use books up to 2008, which a) means that they're happy to use books that aren't just public domain by virtue of age, b) might give us some indication of where these books come from." Nelson Decl. Ex. 344 at -614_001.

476. Discussing the same Google paper, Nick Ryder asked Shantanu Jain, an OpenAI employee who worked on data issues, "if you could pick two things they did we should look into what would it be." Nelson Decl. Ex. 224 at -698.

477. Jain responded: "easy: just magically have their sourcing. they have 600B tokens of books!!!" because OpenAI was "not close to getting that much book data (that we're comfortable training on)." *Id.*

478. Qiming Yuan, an OpenAI technical staff member, added that "Google books [h]as 10 million books that are public[ly] available . . . which means ~1T tokens," though "scraping seems non-trivial and they probably get them directly from Google." *Id.*

479. In June 2023, OpenAI researcher John Schulman wrote to Peter Hoeschele, OpenAI's Vice President of Strategy and Operations for Infrastructure, that he was "pumped about mid-training and starting to acquire high-quality data." Nelson Decl. Ex. 279.

480. When Hoeschele asked whether there were "specific sources you're very excited about," Schulman identified academic publications, books, and Substack, and stated that "books are generally higher quality than stuff on the internet." *Id.*

481. Schulman added that if OpenAI could improve "long context perf[ormance] in mid training" books would be valuable "for that reason." *Id.*

482. In November 2023, while discussing the performance of a competing model, Ryder stated that "books" may explain that model's performance. Nelson Decl. Ex. 236 at -315.

483. Alec Radford, an OpenAI member of technical staff, responded that there was "evidence that books look higher quality on average and that's with the ones we can use right now," and added that it "would be nice to have an expensive but high quality books dataset to de-risk on [as] even [redacted] is enough for data models to get clear signals on[.] [S]o I'd be really supportive of spending [redacted] on a preliminary books sourcing dataset that's high quality and broad coverage." *Id.*

484. In February 2024, Brockman wrote that "as we get better at synthetic data approaches, data sources that you'd intuitively guess a person should be trained on (i.e. books, university coursework/notes, etc) will become the most valuable." Nelson Decl. Ex. 356 at -585.

485. Brockman further wrote that he expected "a data amplifier, where high-quality data is amplified by 1000X or 10,000X in terms of value," and that "in that world, we'd still want lots of high-quality human data—and things like 'books' are probably more valuable than ever." *Id.* at -580.

486. The understanding that books were valuable training data was widely shared in the LLM industry, including among OpenAI's competitors Anthropic and Google, and among third-party researchers. Zhao Decl. Ex. A at ¶¶ 46–72.

487. When describing the value of receiving an additional 1T tokens, an Anthropic researcher stated that "books represent very high quality pretraining data." Nelson Decl. Ex. 740.

488. Anthropic researchers found that long-context (including books) are crucial to an LLM's ability to produce high-quality text "rather than low-quality internet speak." Nelson Decl. Ex. 741, at -433.

489. This is what Anthropic described as "The Good Writing Hypothesis," which amounts to the idea that in order to be effective "[m]odels just need good writing" in their training data. *Id.* at -433.

490. Anthropic's financial forecast described books as part of how it would compete with OpenAI. Anthropic stated that it could compete by "continu[ing] to release the same products as OpenAI . . . but offer them to customers at a lower price." Nelson Decl. Ex. 742.

491. When describing how it would effectuate this strategy, Anthropic stated that data quality would be a "key contributor." *Id.*

492. As Anthropic described it "[a]dding higher quality data (particularly books that aren't publicly available on the internet) allows smaller and cheaper models to outperform larger models." *Id.*

493. Books were listed first on Anthropic's list of sources of additional "High-Quality Data." Nelson Decl. Ex. 741.

494. Members of the strategic finance team stated that "[h]eading into 2025, we continue to have the strongest conviction around book data. Across all categories of written data, books are both more useful and available in larger quantities." Nelson Decl. Ex. 743.

B. <u>Defendants' Experts Acknowledge That Books Were Not Necessary</u>

495. OpenAI's contemporaneous documents establish that it regarded books as among the highest-quality training data available and pursued them for that reason. *See supra* Section IV.A. Defendants' experts do not contend that books failed to improve model performance. Nelson Decl. Ex. 751 ¶¶ 328–35; Nelson Decl. Ex. 727 ¶ 65.

496. Defendants' technical and economic experts opine instead that Defendants did not need the Asserted Works to create "general-purpose" large language models, that no individual work and no category of works including books was indispensable to those models' capabilities,

and that the Asserted Works were therefore "not technically required." Nelson Decl. Ex. 751 ¶ 328; Nelson Decl. Ex. 753 ¶¶ 21, 86, 87; Nelson Decl. Ex. 727 ¶¶ 5, 66; Nelson Decl. Ex. 754 ¶ 98.

497. OpenAI's technical expert, Professor Chris Callison-Burch, opines that "[n]o individual work or set of works is required to achieve a diverse distribution of data," and that "any individual source, particularly in a given domain, is entirely fungible." Nelson Decl. Ex. 751 ¶ 328.

498. Callison-Burch states that Plaintiffs' works "constituted a negligible fraction of the total training data," and that there is "no reason to believe that the expression contained in those works . . . had a meaningful impact on the performance of OpenAI's models." *Id.* ¶ 11(c).

499. He further states that "the inclusion or exclusion of any individual document in the model's training data should not have a meaningful impact on how the model generates answers to any given user query," and that success on OpenAI's goal of having diverse training datasets "in no way depends on the inclusion of any specific document or group of documents in the training set." *Id.* ¶¶ 343–44.

500. Callison-Burch states in reply that "[t]here is no reason to believe that the Asserted Works individually had any meaningful impact on model performance," and that "no individual work was necessary for inclusion in the corpora." Nelson Decl. Ex. 756, ¶ 20(a).

501. Callison-Burch further agreed at his deposition that "[t]he technical necessity of obtaining a digital copy does not establish the necessity of obtaining that copy from LibGen." Nelson Decl. Ex. 11 at 512:17-21.

502. Microsoft's technical expert, John Lafferty, opines that an LLM's general-purpose capabilities are "based upon the statistical associations and patterns of co-occurrence that appear

across the full distribution of the training data, rather than being based in any single document or small collection of texts," and that the training process "is not dependent upon the content of any particular text." Nelson Decl. Ex. 753, ¶ 14.

503. Lafferty states that "no individual work or subset of works is responsible for conferring" an LLM's capacity to generalize. *Id.* ¶ 21.

504. He further states that "no individual text or category of texts within a training corpus of this scale is necessary to the model's capability," and that "the removal of any single work or collection of works from a corpus of trillions of tokens would not materially alter the statistical structure the model learns." *Id.* ¶ 87.

505. Lafferty further opines that "news and book content are not uniquely necessary to LLM capability." Nelson Decl. Ex. 727 at ¶ 5.

506. Lafferty states that "[e]stablishing that news or book text can improve some outputs does not establish that Plaintiffs' particular works were necessary to the accused models' capability, that they made a measurable marginal contribution beyond the many other sources sharing the same attributes, or that any such contribution is attributable to those works[.]" *Id.* ¶ 59.

507. He states that "books may be useful as a training signal since they offer long-form coherence, editorial quality, linguistic diversity, and domain-specific exposition," but that "this does not mean that commercially published copyrighted books are uniquely necessary to general LLM capability." *Id.* ¶ 65.

508. He adds that "[o]ther high-quality sources can also contribute to model quality, and an LLM learns from statistical patterns across the full training distribution, not from any single category of works in isolation." *Id.*

509. Lafferty states that "no individual text or category of texts is necessary to an LLM's capability, and the utility of an LLM does not come from preserving the content of any subset or category of texts." *Id.* ¶ 66.

510. He further states that "[a] model's ability to produce coherent long-form output does not show that any particular category of commercially published books, much less any individual book, is required for general LLM capability." *Id.* ¶ 73.

511. Microsoft's economic expert, Catherine Tucker, opines that "[n]o one piece of data is essential to training a general-purpose LLM because the goal of training is to develop the LLM's ability to generalize based on patterns in data that are derived from the full distribution of the training dataset," such that "the incremental value of adding any one piece of data . . . is close to zero." Nelson Decl. Ex. 754, ¶ 98.

512. Tucker reports that Microsoft's Jordan Usdan testified that tens of thousands or even 100,000 books are "like a drop in the ocean," and that OpenAI's Peter Hoeschele testified that "any individual dataset is kind of vanishingly not useful" and that "any individual piece of data is so miniscule in the larger dataset." *Id.*

513. Tucker also reports that OpenAI's Nicholas Ryder testified that books were only "a very small percentage" of OpenAI's training data. *Id.* ¶ 100. Tucker also opines that the 211 Asserted Works were a "miniscule share" of the relevant datasets. *Id.* Tucker observed that Microsoft's Jordan Usdan stated, "there's been no evidence that [the Harper Collins] books made a difference in helping [Microsoft's] models." *Id.* ¶ 133 (second alteration in original).

514. OpenAI and Microsoft fact witnesses gave similar testimony. Nicholas Ryder testified that there were many investigations into what constituted high-quality text and that, "for

the purposes of pretraining, . . . the most important factor by far is getting rid of spam, getting rid of SEO, getting rid of obscene material." Nelson Decl. Ex. 50(a) at 33:3–6.

515. Ilya Sutskever likewise testified that, while books might be "useful to some degree" for a function like semantic search, such an application "is not something that requires the books." Nelson Decl. Ex. 58 at 101:20–102:11. For such a feature, "you could definitely do it without" books. *Id.* at 102:9–11.

516. Microsoft employee Jordan Usdan testified that there was "no evidence" that the books it received from [redacted] "made a difference in helping [Microsoft's] models" and that it would be hard to see the difference from "even a hundred thousand books." Nelson Decl. Ex. 63(b) at 78:21–79:18.

C. <u>OpenAI Used LibGen to Train GPT-3 and GPT 3.5</u>

517. In April 2019, OpenAI created a training dataset from the files Radford downloaded from LibGen. *See supra* § II.B; *see also* Nelson Decl. Ex. 234 at -960; Shan Decl. Ex. A at ¶ 50. OpenAI first called that dataset "LibGen" and later designated it "LibGen1." Shan Decl. Ex. A at ¶ 50.

518. In December 2019, OpenAI created a training dataset from the files Mann torrented from LibGen. *See supra* § II.E. Shan Decl. Ex. A at ¶ 62. OpenAI called that dataset "LibGen2." Nelson Decl. Ex. 33 at 93:18–94:1; Shan Decl. Ex. A at ¶ 62.

519. OpenAI used LibGen1 and LibGen2 to train GPT-3 and GPT-3.5, including an early version of GPT-3 that it demonstrated to Bill Gates and other senior Microsoft executives in April 2019. Nelson Decl. Ex. 248 at -671; Nelson Decl. Ex. 337; Nelson Decl. Ex. 62(a) at 52:1–14; Nelson Decl. Ex. 316.

520. Testifying as OpenAI's corporate designee, Alex Paino confirmed that "LibGen-1 and LibGen-2 were used to train GPT-3 described in the public paper." Nelson Decl. Ex. 43(a) at 234:20–24.

521. OpenAI trained the GPT-3 models between March and April 2020 on a data mix called [redacted]. Nelson Decl. Ex. 316; Nelson Decl. Ex. 757 at -001.

522. In May 2021, [redacted] was a [redacted] token mixture used in [OpenAI's] largest runs," containing [redacted] tokens from LibGen1 [redacted] tokens from LibGen2. Nelson Decl. Ex. 316.

523. Shantanu Jain wrote that "GPT3 was trained on the equivalent of [redacted] books (if you extrapolate based on the size of the books in the current training mix)." Nelson Decl. Ex. 349 at -095.

524. Dario Amodei testified that the as released version of GPT-3 "included LibGen." Nelson Decl. Ex. 4 at 109:3–6. OpenAI used a version of LibGen1 titled "libgen1-dedup-clean" and a version of LibGen2 titled "libgen2-dedup-clean" to train GPT-3. Nelson Decl. Ex. 316; Shan Decl. Ex. A ¶¶ 51, 62.

525. OpenAI located and produced a copy of "libgen1-dedup-clean" at OPCO_SDNY_1457402, and a copy of "libgen2-dedup-clean" at OPCO_SDNY_1457403. Shan Decl. Ex. A ¶¶ 50–51, 63.

526. OpenAI trained the GPT-3.5 models between October and December 2021 on a data mix known as [redacted]. Nelson Decl. Exs. 316, 338, 340.

527. Testifying as OpenAI's corporate designee, Ian Sohl confirmed [redacted] was the data mix used for GPT-3.5. Nelson Decl. Ex. 57(a) at 448:16–20.

528. A September 2021 document describing the components of [redacted] lists "libgen1-dedup-clean-csam-filt" at [redacted] tokens, and records that "[s]exual content" had been "removed." Nelson Decl. Ex. 369 at -937.

529. OpenAI used that dataset and two versions of LibGen2, titled "libgen2-enonly-dedup-clean-csam-filt" and "libgen2-nonen-dedup," to train GPT-3.5. Nelson Decl. Ex. 316; Shan Decl. Ex. A ¶¶ 51, 62.

530. Paino testified that "some variants of LibGen-1 and LibGen-2 were used to train GPT-3.5," and that those variants reflected "some additional filtering that occurred relative to the GPT-3 version of these datasets." Nelson Decl. Ex. 43(a) at 235:4–12.

531. OpenAI has never located those three GPT-3.5 datasets following its deletion of the LibGen data in the summer of 2022. Shan Decl. Ex. A ¶¶ 51, 63. OpenAI states that they were filtered versions of "libgen1-dedup-clean" and "libgen2-dedup-clean," such that the books present in the produced datasets would also have been present in the datasets used to train GPT-3.5. Shan Decl. Ex. A ¶¶ 51, 63.

D. OpenAI Also Wanted to Use Books3 to Train Its Models

532. OpenAI also trained models on The Pile, which includes the Books3 component assembled from the shadow library Bibliotik. Nelson Decl. Ex. 37 at 36:10–38:10; Nelson Decl. Ex. 277.

E. OpenAI Also Trained the Accused Models on Datasets Derived from Its Scrapes and Crawls

533. OpenAI also copied the books contained in its targeted scrapes and third-party crawls into the datasets it used to train the Accused Models. OpenAI trained every Accused Model other than GPT-4o mini on datasets derived from Common Crawl, and trained GPT-4o on datasets

derived from the GPT-Bot and Project Mango crawls. Shan Decl. Ex. A ¶¶ 117–19, 134–38; Shan Decl. Ex. H; Shan Decl. Ex. I; Shan Decl. Ex. J.

534. GPT-4o mini did not train on any data containing the Asserted Works, because OpenAI trained it [redacted]. Shan Decl. Ex. B ¶ 28. GPT-4o mini is nonetheless a performant and functional large language model. Nelson Decl. Ex. 751 ¶ 263.

F. <u>OpenAI Created Multiple Copies of Books in Training</u>

535. Training a large language model requires making multiple copies of the data at each stage of the training pipeline. Nelson Decl. Ex. 215 at -233; Nelson Decl. Ex. 43(a) at 302:11-304:3.

536. OpenAI creates a copy when it converts raw text files into training datasets. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Ex. 43(b) at 146:24-148:22; Nelson Decl. Ex. 208 at -516.

537. OpenAI creates an additional copy when it filters training datasets. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Ex. 43(b) at 146:24-148:22.

538. OpenAI creates an additional copy when it samples training datasets. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Ex. 43(b) at 146:24–148:22.

539. OpenAI creates an additional copy when it cleans training datasets. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Ex. 43(b) at 146:24–148:22.

540. OpenAI creates an additional copy when it tokenizes the data. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Nelson Decl. Ex. 43(b) at 146:24–148:22; Nelson Decl. Ex. 208 at -516.

541. OpenAI creates an additional copy when it shards the tokenized data. Shan Decl. Ex. B ¶ 41; Nelson Decl. Ex. 43(b) at 171:1–10.

542. OpenAI creates an additional copy when it caches the data. Shan Decl. Ex. B ¶ 41.

543. OpenAI creates an additional copy when it loads the dataset into the model training environment. Shan Decl. Ex. A ¶¶ 70–71; Nelson Decl. Ex. 43(b) at 146:24–148:22; Nelson Decl. Ex. 208 at -516.

544. Copying repeats each time OpenAI passes the dataset through the model during a training epoch. Shan Decl. Ex. B ¶¶ 41–42.

545. A June 2022 inventory of OpenAI's storage lists pretokenized copies of the LibGen datasets in multiple directories, including "pretokenized/reversible_50000/libgen-1-dedup-clean," "pretokenized/programming_v1/libgen1-dedup-clean-csam-filt," and "pretokenized/distaug_programming_v1/libgen2-enonly-dedup-clean-csam-filt." Nelson Decl. Ex. 340 at -001, -002, -016.

546. Before OpenAI trained [redacted], OpenAI employee Lilian Weng ran both LibGen datasets through a filter to exclude certain content, creating numerous further copies that she stored on her OpenAI Azure personal container [redacted] and named, among others, "[redacted]-data-cleanup/filtered/libgen-1," "[redacted]-data-cleanup/filtered/libgen-1-dedup," "[redacted]-data-cleanup/libgen-2," [redacted]-data-cleanup/filtered/libgen-2-dedup," "[redacted]-data-cleanup/libgen1-v3," and "[redacted]-data-cleanup/libgen2-v3." Shan Decl. Ex. A ¶ 70.

547. Testifying as OpenAI's corporate representative, Paino stated that it would be difficult to count how many times copying occurs during the process of training a large language model. Nelson Decl. Ex. 43(a) at 302:11–303:6.

G. <u>The Accused Models Reproduce Class Plaintiffs' Works to End Users</u>

548. The Accused Models have the capability to, and in fact do, reproduce portions of the Asserted Works to end users when prompted, including in response to requests for summaries of the Asserted Works. Shan Decl. Ex. A ¶¶ 145–49, 154; Shan Decl. Ex. K; Shan Decl. Ex. B ¶ 45.

549. Plaintiffs' expert Dr. Shawn Shan analyzed the output logs Defendants produced in this litigation and documented actual reproduction of the Asserted Works to end users. Dr. Shan treated a model-authored message as a regurgitation of a Plaintiff's work only if its longest contiguous near-verbatim block was at least thirty consecutive words, and he excluded from each model message any text echoing the user's own prompts. Shan Decl. Ex A ¶¶ 38–41.

550. Applying that methodology to OpenAI's output logs, Dr. Shan identified reproductions of thirty or more near-verbatim words from seven of Plaintiffs' Works across thirty-one conversations and thirty-three chat responses. Shan Decl. Ex A ¶ 154; Shan Decl. Ex. K.

551. Applying the same methodology to Microsoft's 8.2 million Bing Chat output logs, he identified reproductions of thirty or more near-verbatim words from ten of Plaintiffs' Works across twenty-four conversations and twenty-four chat responses. Shan Decl. Ex A ¶¶ 155–56; Shan Decl. Ex. K.

552. Those counts are a lower bound. The output log samples OpenAI produced generally do not encompass the earlier periods when regurgitation would have been more prevalent, and represent only a fraction of the overall output logs. Shan Decl. Ex A ¶¶ 151–53.

553. An OpenAI employee running regurgitation studies observed at the time the 78 million-log sample was collected that it would be "a little light given how low our hit rate." Nelson Decl. Ex. 294 at -468.

554. The redactions applied to the produced logs also obscure the names of characters from Plaintiffs' Works, impeding identification of infringing outputs. Shan Decl. Ex A ¶ 152.

555. Reproduction also occurs in response to ordinary, non-adversarial requests of the kind users make every day, including requests for summaries. A one-sentence request for a detailed summary of a named Asserted Work by a named Plaintiff author produced multi-paragraph, plot-

faithful descriptions averaging approximately 400 words and reaching a maximum of 2,353 words, discussing named characters, key plot points, and how the story concludes. Shan Decl. Ex. B ¶¶ 7, 53, 55.

556. Six of those summary outputs contained verbatim passages of thirty or more words from the named Plaintiffs' books, even though the prompts requested only a summary and supplied no text from any book. Shan Decl. Ex. B ¶¶ 7, 22, 31.

557. In 2022, OpenAI started Project Giraffe, whose goal was to "develop techniques to further limit regurgitation" of training data. Nelson Decl. Ex. 38(a) at 35:13–18. The Giraffe filter was intended to detect and block regurgitation of certain copyrighted material identified on certain blocklists. Nelson Decl. Ex. 360 ("To prevent copyright regurgitation, we have deployed mitigations at various stages. SafeSys mainly worked on two mitigations under the code name 'Giraffe': inference-time bloom filter (BF) blocking and post-training intervention."); Nelson Decl. Ex. 359 at -436 ("At the inference time, we deployed a distributed, large-scale bloom filter to remove n-gram phrases that match any known copyrighted materials in our database.").

558. Prior to the release of ChatGPT, OpenAI was aware that its models could "memorize and later regurgitate training data under certain circumstances." Nelson Decl. Ex. 38(b) at 260:23–261:4.

559. OpenAI personnel described memorization not as an accident but as a product of the training objective itself, and understood that memorization was positively correlated with their own evaluations of model quality. An internal OpenAI presentation states that "models want to regurgitate, so we're inherently fighting against the training objective," and that "this is the first time we are trying to make LLMs bad at something." Nelson Decl. Ex. 298 at -468; Shan Decl. Ex. A ¶ 142.

560. Another OpenAI document states that "[l]arge overparametrized neural networks can see a book a handful of times and they'll just memorize it . . . . We train our networks to memorize the training data—that's their objective . . . our networks have lots of parameters and we train them to memorize the dataset." Nelson Decl. Ex. 274 at -365–66; Shan Decl. Ex. A ¶ 142.

561. A third states: "Looking forward, though, I think we should assume future big models will be able to memorize everything in the train corpus, even if it only appears once per epoch." Nelson Decl. Ex. 286; Shan Decl. Ex. A ¶ 142.

562. When attempting to address the problem, OpenAI employees described it as "a fundamental contradiction for our models to seek to memorize text and recall it verbatim, except in certain arbitrary instances." Shan Decl. Ex. B ¶ 45.

563. In March 2022, in a document titled "Update on Preventing Regurgitation," OpenAI recognized that a "major problem" with GPT-3.5 is that it "can easily reproduce training data verbatim." Nelson Decl. Ex. 292.

564. OpenAI released GPT-3, GPT-3.5, and ChatGPT without any anti-regurgitation protections, and it continued to update ChatGPT with new models while recognizing internally that regurgitation had not been solved. Shan Decl. Ex. A. ¶¶ 140, 142–44.

565. OpenAI did not deploy the initial version of Giraffe filter to ChatGPT until a few weeks after ChatGPT's November 30, 2022 launch. Nelson Decl. Ex. 38(b) at 212:6–16.

566. The first version of the Giraffe filter was an "in-memory" filter and included a limited number of copyrighted materials such as Harry Potter, the famous "Quake inverse square root" code snippet, and snippets from approximately 20 books. Nelson Decl. Ex. 38(b) at 299:9–300:14.

567. Plaintiffs' Works are present throughout the files used in and generated as part of Project Giraffe. The files "[redacted]" reflect forty-six of Plaintiffs' Works by nine Plaintiffs. Shan Decl. Ex. A. ¶ 145.

568. When the Giraffe team ran internal searches for authors and book titles, it located ninety-six of Plaintiffs' Works. Shan Decl. Ex. A ¶ 145.

569. OpenAI's own evaluations run on Project Giraffe show that the models regurgitated Plaintiffs' Works: the file "[redacted]" reflects regurgitation of five Asserted Works by three Plaintiffs, and the file [redacted] lists eight Asserted Works by six Plaintiffs. Shan Decl. Ex. A ¶¶ 146–47.

570. Taken together, the identification of Plaintiffs' Works in the Project Giraffe evaluations, OpenAI's own findings that its training data could be easily memorized by its models, and the regurgitations documented in Defendants' output logs establish that the Accused Models memorized at least portions of Plaintiffs' Works from their training data and stored those portions inside the model. Shan Decl. Ex. A ¶ 150.

V. OpenAI and Microsoft Distributed Books to Each Other and Third Parties

571. In the course of their commercial relationship, OpenAI and Microsoft exchanged copies of the LibGen data with each other. Shan Decl. Ex. A. ¶ 72.

572. OpenAI also distributed LibGen data to third-party contractors. Shan Decl. Ex. A ¶ 72; Nelson Decl. Ex. 42 at 111:4–112:16.

A. OpenAI Provided Microsoft Access to GPT-3 Training Data

573. Under a September 2020 amendment to the 2019 OpenAI-Microsoft agreement, OpenAI agreed to provide [redacted]

[redacted] Nelson Decl. Ex. 158 at -967.

574. OpenAI's corporate witness Benton Gaffney agreed that this provision entitled [redacted]. Nelson Decl. Ex. 18 at 145:14–146:4.

[redacted]

[redacted]

[redacted]

578. OpenAI also shared portions of the GPT-3 training data with Microsoft on other occasions. Nelson Decl. Ex. 41 at 46:15–47:11; Nelson Decl. Ex. 275 at -591; Shan Decl. Ex. A ¶ 72(d).

579. In November 2021, Microsoft shared access to GPT-3 training datasets [redacted] Nelson Decl. Ex. 1 at 82:4–20; Nelson Decl. Ex. 780. Microsoft Transferred the Bing Index to OpenAI (Project Taxi). Shan Decl. Ex. A ¶ 127.

580. From the outset of its relationship with Microsoft, OpenAI sought a copy of the "Bing Index," the internet-wide crawl of the web that Microsoft built for its Bing search engine. Nelson Decl. Ex. 163 at -609; Nelson Decl. Ex. 159 at -571; Shan Decl. Ex. A ¶ 126.

581. On November 6, 2019, the parties entered into [redacted], under which Microsoft agreed to make available to OpenAI certain data from Microsoft's Bing Index, described as [redacted] Nelson Decl. Ex. 161 at –041–42.

582. On March 25, 2020, the parties amended the statement of work to provide that OpenAI's use of the Bing Index data would be for "computational analysis," defined to include [redacted] Nelson Decl. Ex. 178.

583. Common Crawl was a "subsample of the web," Nelson Decl. Ex. 50(a) at 159:8–13, while the Bing Index contained "[redacted]." Nelson Decl. Ex. 63(b) at 15:9–20.

584. The Bing Index crawler can also access paywalled and copyrighted content that web publishers may restrict from general crawlers. Nelson Decl. Ex. 63(a) at 138:11–140:4.

585. [redacted] Nelson Decl. Ex. 143 at -215-216. OpenAI continued to press for the data through 2021 and 2022. Nelson Decl. Ex. 19(a) at 111:1–119:9.

[redacted]

587. Following negotiations between OpenAI's Chief Technology Officer, Mira Murati, and Microsoft, the parties agreed [redacted] *Id.*

588. OpenAI employee Bob McGrew described the arrangement as follows: "the idea [was] that we're giving data to them and they are giving data to us, [redacted] Nelson Decl. Ex. 779.

589. Microsoft conditioned the Bing Index transfer on OpenAI's provision of training data. Microsoft employee Karmel Allison, who proposed to "[redacted], called the arrangement "horse trading." Nelson Decl. Ex. 1 at 66:16–69:24 ; Nelson Decl. Ex. 173 at -440; Nelson Decl. Ex. 174; Lasinski Decl. Ex. A ¶ 28. Allison testified that the exchange "is about a conversation and horse trading." Nelson Decl. Ex. 1 at 69:11–24.

590. Katherine Mayer testified that OpenAI [redacted] because it was [redacted] Nelson Decl. Ex. 826 at 163:13–25.

591. OpenAI [redacted] Nelson Decl. Ex. 781 at -435 (emphasis added [redacted] Nelson Decl. Ex. 173.

[redacted].

593. After receiving the data it sought from OpenAI, Microsoft transferred full copies of the Bing Index to OpenAI in an effort called "Project Taxi." Nelson Decl. Ex. 57(a) at 344:16–24; Nelson Decl. Ex. 67 at 77:18–21; Shan Decl. Ex. A ¶ 127.

[redacted]

[redacted]

595. Ribas, who is "the head of search at Microsoft," testified that he was not "aware of any restrictions at all on what OpenAI could do with the Bing index data when it was transferred to OpenAI." Nelson Decl. Ex. 47 at 17:9–16; 199:11–16.

596. OpenAI employees who trained models on the Bing Index data reported that it performed as well as OpenAI's prior Common Crawl web crawl. Nelson Decl. Ex. 67 at 80:3–6; Shan Decl. Ex. A ¶ 128. OpenAI deleted the Bing Index data in late 2023. Nelson Decl. Ex. 57(a) at 346:6–13; Nelson Decl. Ex. 170.

B. <u>Microsoft Transferred Crawled Data to OpenAI (Project Mango)</u>

597. After Microsoft declined to permit OpenAI to use Bing Index data to train its publicly available models, OpenAI decided to conduct its own crawl of the internet. Nelson Decl. Ex. 19(a) at 142:23–143:24.

598. Microsoft transferred to OpenAI a list of URLs to guide OpenAI's crawler. Nelson Decl. Ex. 782; Nelson Decl. Ex. 12 at 111:1–113:19; Nelson Decl. Ex. 19(b) at 36:6–37:15, 188:22–190:10; Shan Decl. Ex. A ¶ 134.

599. OpenAI lacked the crawling infrastructure and search-engine experience Microsoft had developed for Bing [redacted] Nelson Decl. Ex. 19(b) at 214:19–215:14; Nelson Decl. Ex. 60 at 184:7–185:10; Nelson Decl. Ex. 170.

600. OpenAI then signed a [redacted] contract with Microsoft under which OpenAI would crawl half of the web on its own servers and Microsoft would crawl the other half and transfer the downloaded content to OpenAI. Nelson Decl. Ex. 783; Nelson Decl. Ex. 57(a) at 336:20-344:1; Shan Decl. Ex. A ¶ 135.

601. The parties called this effort "Project Mango," under which Microsoft crawled the internet to "collect data for the purposes of OpenAI to train LLM models." Nelson Decl. Ex. 19(b) at 140:5-18; Nelson Decl. Ex. 12 at 71:10–72:11; Nelson Decl. Ex. 57(a) at 337:17-24.

602. In August and September 2023, OpenAI decided to crawl the full internet using a program called "GPT-Bot," and the resulting data was known within OpenAI as the "GPT-bot" data. Nelson Decl. Ex. 57(a) at 456:8–16.

603. Dr. Shan testified that, for the Project Mango crawling, Microsoft was to use OpenAI's user agent, "GPTBot," and that to his knowledge Microsoft used a different crawler—which he understood may have been called "Bing bots"—for its ordinary crawling of the Bing Index. Nelson Decl. Ex. 53 at 309:2–24; Shan Decl. Ex. A ¶ 135; Nelson Decl. Ex. 57(a) at 337:17–24.

604. Microsoft began the Project Mango crawling service in August 2023 and began transferring the crawled data to OpenAI in September 2023. Lasinski Decl. Ex. A ¶ 91. The list of "50k top domains to be crawled by MSFT" drawn from the Bing Index included the LibGen domain "libgen.ee," as well as thepiratebay.ee, zlibrary.cc, and e-books.com. *Id.*; Nelson Decl. Ex. 12 at 106:1-108:18.

605. Microsoft's corporate designee Fabrice Canel testified that Microsoft did not check whether the data it crawled from those domains and transferred to OpenAI contained copyrighted works, and that Microsoft did not "care" if there were copyrighted books in that data. Asked whether there were "any processes at all at Microsoft to determine whether there are copyrighted works in the Bing index data that they downloaded or that they shared with third parties," Canel answered: "No." Asked whether Microsoft "care[s] whether there is copyrighted works and

copyrighted books in Bing index data that it shares with third parties," Canel again answered: "No." Nelson Decl. Ex. 12 at 37:21–38:9.

606. [redacted] Nelson Decl. Ex. 12 at 84:1–3. At least one Microsoft employee objected to Microsoft's role in Project Mango. Nelson Decl. Ex. 12 at 77:3–78:9.

C. <u>OpenAI Transferred LibGen to Book-Summarization Contractors</u>

607. OpenAI distributed copies of the LibGen books to third-party contractors as part of its book-summarization project, which OpenAI undertook to improve its models' performance at "instruction following"—that is, obeying users' commands. Nelson Decl. Ex. 785 at 21:4–23:19, 110:20-111:6; Nelson Decl. Ex. 42 at 111:4–112:16; *see generally* Nelson Decl. Ex. 787.

608. OpenAI drew on the LibGen books it retained in its files and served them to the contractors in chunks to read and summarize. Nelson Decl. Ex. 785 at 110:7–111:6.

609. Jeff Wu, the researcher primarily responsible for the LibGen portions of the project, understood that using LibGen materials violated copyright restrictions. Nelson Decl. Ex. 785 at 86:4–87:6; Nelson Decl. Ex. 788.

610. Although OpenAI deleted the full set of files used in this project, Wu retained a store of LibGen books in his personal files under the name "booksummaries." Shan Decl. Ex. A ¶ 72.

611. OpenAI has no records reflecting whether the third-party contractors retained the LibGen books they received or what else they did with those books. Nelson Decl. Ex. 42 at 185:14–186:15.

VI. Microsoft's Role in OpenAI's Infringement

A. OpenAI and Microsoft Raced to Close Google's Lead in Compute, Talent, and Books

612. By 2018, Google had emerged as the leading company in artificial intelligence. Nelson Decl. Ex. 105.

613. OpenAI's research showed promise, but the company employed few people and had little money to buy the data and compute its technology needed. *See* Nelson Decl. Ex. 2(a) at 25:7–26:2; Nelson Decl. Ex. 32 at 91:21–92:9.

614. Microsoft faced the opposite constraint: it held more resources and compute than most companies, but its research and innovation lagged behind OpenAI's and Google's. Nelson Decl. Ex. 40 at 202:4–25, 212:14–18; 247:20–248:12; Nelson Decl. Ex. 682 at 1; Nelson Decl. Ex. 152 at -738–39; Nelson Decl. Ex. 60 at 366:16–368:3; Nelson Decl. Ex. 157 at -364; Nelson Decl. Ex. 160 at -904.

615. In 2016, Google publicly declared itself an "AI First" company. Nelson Decl. Ex. 75 at 4.

616. In June 2017, Google researchers published "Attention Is All You Need," introducing the transformer architecture on which OpenAI later built its GPT models. Nelson Decl. Ex. 92; Nelson Decl. Ex. 789 at 4.

617. Satya Nadella, Microsoft's CEO, testified that as of 2018, Google "not only has all the strengths in distribution, all the – which is the search, YouTube, they have all the data, again, search, YouTube, they have all the training, compute and talent and the TPUs, and so for us, it has been clear that since then, that unless we can somehow make this entire market a little more competitive, none of us would stand a chance." Nelson Decl. Ex. 40 at 247:20–248:12; Nelson Decl. Ex. 152 at -737.

Nadella testified that by 2019, Microsoft suspected Google's AI efforts were "way ahead of" Microsoft's, with a likely "decade long" headstart. Nelson Decl. Ex. 40 at 202:4–25, 212:14–18; Nelson Decl. Ex. 682 at 1.

619. Testifying as Microsoft's corporate designee, Jon Tinter stated that Google "had some inherent advantages," including "the interrelationships to search, and a lot of the technology that developed in search, the systems and scale expertise, as well as they started investing in this early and just the absolute level of investment that they had." Nelson Decl. Ex. 60 at 366:16–368:3.

620. Tinter further testified: "A lot of the foundational research around the entire GPT transformer technology came out of Google research, you know, so they had significant talent. They had a first mover advantage. They had assets in terms of their systems capability. They had assets in terms of the data they had on search." Nelson Decl. Ex. 60 at 366:16–368:3.

624. Before the Microsoft partnership, OpenAI used Google's cloud platform for computing. Nelson Decl. Ex. 40 at 248:13–25; Nelson Decl. Ex. 166 at -441. Nadella testified that OpenAI "were coming to the realization that this is no longer about some free credits, but you needed a real commercial partner and a long-term construct," and that he "assum[ed] Google was not interested because they had their own effort." Nelson Decl. Ex. 40 at 248:13–25.

625. An OpenAI presentation on the Microsoft partnership listed as its first goal: Nelson Decl. Ex. 790 at Slide 2. The presentation stated that

[redacted]" *Id.* at Slide 4.

626. On April 18, 2019, Dario Amodei, OpenAI's then-Research Director, wrote that [redacted]" Nelson Decl. Ex. 287 at -399. Amodei stated that OpenAI's message should be: [redacted]" *Id.*

[redacted]

628. Google's lead extended to books. Beginning in 2004, Google spent more than a decade building a digital database of approximately 40 million scanned books. *Authors Guild v. Google, Inc.*, 804 F.3d 202, 208 (2d Cir. 2015); Nelson Decl. Ex. 76.

[redacted]

B. <u>Microsoft Formed a Partnership with OpenAI</u>

[redacted]

[redacted].

632. In February 2019, OpenAI published "Language Models are Unsupervised Multitask Learners," describing GPT-2. Nelson Decl. Ex. 789. The paper's publication led to a meeting between Sam Altman and Satya Nadella, followed by the April 2019 demonstration for Bill Gates at which OpenAI disclosed its use of LibGen. *See supra* § II.D. Nelson Decl. Ex. 182 at -073, -085.

[redacted]

634. In July 2019, Microsoft agreed to invest $1 billion in OpenAI. Nelson Decl. Ex. 147 at Slide 5; Nelson Decl. Ex. 203 at 855; Nelson Decl. Ex. 30 at 38:23–39:6. In 2021, Microsoft invested another $2 billion. Nelson Decl. Ex. 147 at Slide 5; Nelson Decl. Ex. 203 at 855. In 2023, Microsoft committed an additional $10 billion. Nelson Decl. Ex. 203 at 855.

635. Microsoft's total capital commitments to OpenAI amounted to at least $13 billion. Nelson Decl. Ex. 271 at -570–71 (Note 13 to OpenAI Global, LLC Consolidated Financial Statements for the Year Ended December 31, 2023).

[redacted]

637. On July 2, 2019, Microsoft and OpenAI entered into a Joint Development and Collaboration Agreement ("JDCA") to collaborate on designing and building a supercomputer and

supercomputer processing unit. Nelson Decl. Ex. 145. Microsoft agreed to provide OpenAI with dedicated Azure hardware. *Id.* at -354, -359.

638. Microsoft's $1 billion investment consisted of [redacted] Nelson Decl. Ex. 145 at -359; Nelson Decl. Ex. 203 at 855; Nelson Decl. Ex. 60 at 312:21–314:23.

639. Sam Altman signed the JDCA for OpenAI, Inc. and the new for-profit entity OpenAI, LP; Satya Nadella signed for Microsoft. Nelson Decl. Ex. 145 at -345–47.

[redacted]

[redacted]

[redacted]

646. In March 2021, the parties signed the Amended and Restated JDCA. Nelson Decl. Ex. 150 at -084-085.

649. In January 2023, the parties signed the Second Amended and Restated JDCA ("Second Amended JDCA"). Nadella signed for Microsoft; Altman signed for OpenAI, Inc. and OpenAI OpCo, LLC (successor to OpenAI LP). Nelson Decl. Ex. 146 at -575–77. That agreement also entitled Microsoft to a share of OpenAI's revenues. *See infra* § VI.G.

[redacted]

[redacted]

[redacted]

C. Microsoft Knew OpenAI Used LibGen

653. As described above, OpenAI disclosed its use of LibGen to Microsoft in April 2019, linking the Wikipedia page that described LibGen as providing "free access to content that is otherwise paywalled" and referenced the 2015 SDNY order directing the site to shut down. *See supra* § II.D and *infra* this section. Microsoft's knowledge extended beyond that initial disclosure.

654. In April 2019, Umesh Madan, a Partner and Technical Advisor to Microsoft's Chief Technology Officer, received the OpenAI update stating that OpenAI had added "~11B words from Library Genesis" to its training data. Nelson Decl. Ex. 188 at -958; Nelson Decl. Ex. 189 at -964.

655. The update hyperlinked the Library Genesis entry on Wikipedia, which stated that LibGen provided "free access to content that is otherwise paywalled or not digitized elsewhere" and referenced the 2015 order of the United States District Court for the Southern District of New York directing LibGen to shut down. Nelson Decl. Ex. 189 at -964; Nelson Decl. Ex. 795.

656. Madan, who attended the April 2019 Gates demonstration, testified that OpenAI confirmed it had trained its model on books. Nelson Decl. Ex. 661 at 28:12–29:10, 51:23–52:2.

657. Madan testified that had he known there was "some possibility of piracy" in the book datasets used by his project, "I would have informed our legal department." Nelson Decl. Ex. 32 at 79:13–80:4.

658. In September 2019, Eric Chung shared OpenAI research papers with Microsoft employees including Madan. Nelson Decl. Ex. 191 at -134. One paper authored by Dario Amodei, Alec Radford, and other senior OpenAI researchers ended a paragraph describing training data with the question: "Is LibGen legit?" Nelson Decl. Ex. 192 at -141. The paper provided no answer. *Id.*

659. [redacted] OpenAI's Katie Mayer distributed the same document to Jamie Huynh and Umesh Madan on February 21, 2020. Nelson Decl. Ex. 697. Microsoft's corporate representative admitted that "Microsoft was aware before the release of GPT-3 that OpenAI used Library Genesis or LibGen to train GPT-3." Nelson Decl. Ex. 59(a) at 292:2–17, 296:22–297:2.

660. Mira Murati, OpenAI's CTO, acknowledged that OpenAI "clearly" disclosed its use of LibGen material to Microsoft in that February 2020 document. Nelson Decl. Ex. 39 at 47:21–51:19.

661. Microsoft did not ask OpenAI to withhold GPT-3 from release. Dee Templeton testified that Microsoft "didn't ask OpenAI to stop and not release the GPT-3 model because there was books and novels from Library Genesis in the training data." Nelson Decl. Ex. 59(a) at 299:19–23.

662. Microsoft never removed LibGen from the Bing index. Microsoft's corporate designee testified that Microsoft has "no way to know what is copyrighted versus not" in its index, and that its only mechanisms for excluding content are DMCA notices, robots exclusions, and declining to follow torrents. Nelson Decl. Ex. 19(a) at 107:13–108:4, 150:12–151:5. [redacted]

663. Microsoft publicly states that it does not train Microsoft Copilot "on data from domains listed in the Office of the United States Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy List." Nelson Decl. Ex. 832 at 5; Nelson Decl. Ex. 40 at 185:10–23; Nelson Decl. Ex. 796 at 99; Nelson Decl. Ex. 797 at 90. The Trade Representative has named LibGen in that review since at least 2020. Seeley Decl. Ex. A at ¶ 44; Nelson Decl. Ex. 728 at 27; Nelson Decl. Ex. 40 at 186:4–188:7.

664. Glen Weyl, a Microsoft research lead, expressed concerns about copyright infringement in connection with AI training by OpenAI and Microsoft as early as mid-to-late 2020. Nelson Decl. Ex. 65 at 73:22–77:20.

[redacted]

[redacted]

[redacted]

669. Sarah Bird, Microsoft's corporate designee, testified that "for a technology like this, I think it's very difficult to actually develop or use the technology in any way and bring these risks completely to zero." Nelson Decl. Ex. 6, at 106:5-107:4.

D. Microsoft Built Bespoke Supercomputing Systems for OpenAI's Model Training

671. Lucas Melton, Microsoft's 30(b)(6) designee on the OpenAI supercomputers, testified that Microsoft built and deployed four phases of supercomputer systems for OpenAI, that Microsoft knew OpenAI would use them "to train large language models," and that the systems

were "designed specifically to train OAI's AI models." Nelson Decl. Ex. 36 at 42:11–24, 89:13–24.

673. Nelson Decl. Ex. 27 at 22:16–22:24, 24:4–7; Nelson Decl. Ex. 36 at 39:22–40:9, 41:4–41:11, 57:7–62:25. Melton confirmed that "all GPT models from GPT-3 onward have been trained using Microsoft's Azure supercomputing infrastructure." Nelson Decl. Ex. 36 at 39:22–41:11.

674. OpenAI trained GPT-3 on Phase 1 (Owl) and GPT-3.5 and GPT-3.5 Turbo on Phase 2 (Telemachus). Nelson Decl. Ex. 36 at 39:22–41:11, 84:1–13.

675. That storage was critical to training the at-issue LLMs. Nelson Decl. Ex. 36 at 57:7–62:25; Nelson Decl. Ex. 165 at -548

676. Mark Russinovich, Microsoft's Chief Technology Officer for Azure, testified that he understood OpenAI to be "using Microsoft Azure systems to store pretraining data," and that if that pretraining data contained data from Library Genesis, he understood it to be on Azure's systems. Asked how he would feel about Microsoft systems being used to store that data, Russinovich answered: "I would wonder if it was a violation of our terms of service." Nelson Decl. Ex. 827 at 125:13–126:2.

677. Microsoft also provided Azure supercomputing systems for OpenAI's inferencing operations ([redacted] enabling OpenAI to make models available to users. Nelson Decl. Ex. 36 at 85:14–87:2.

678. When OpenAI makes ChatGPT available to users, it uses Microsoft's Azure systems. Nelson Decl. Ex. 36 at 85:14–87:2.

[redacted]

E. <u>Microsoft Encouraged OpenAI's Data Practices.[redacted] at OpenAI, and Obtained OpenAI's Technology</u>

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

686. Microsoft pursued the collaboration for commercial returns. Nelson Decl. Ex. 44, at 43:1–14; Nelson Decl. Ex. 172, at -643–44. Nadella testified that "from the very beginning" of discussions with Altman and the OpenAI team, Microsoft understood it "would be undertaking this investment for Microsoft's commercial purposes." Nelson Decl. Ex. 40 at 210:11–211:2. Mikhail Parakhin, then a Microsoft Corporate Vice President, testified that as of March 2020 Microsoft aimed to commercialize OpenAI's research to drive $100 million in annual revenue. Nelson Decl. Ex. 44, at 43:1–14.

687. In September 2020, Microsoft discussed internally how to incorporate GPT-3 and other OpenAI products into its commercial offerings, and its need to obtain the GPT-3 training data. Nelson Decl. Ex. 154, at -759–61.

F. <u>Microsoft Could Control OpenAI's Training Data</u>

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

G. Microsoft Benefitted Financially from OpenAI's Infringement

701. Under the Second Amended and Restated JDCA, OpenAI agreed to transfer to Microsoft twenty percent of net revenue generated through commercial use of technology embodying licensed intellectual property. Nelson Decl. Ex. 146 at -596; Nelson Decl. Ex. 23, at 14:25-15:4, 53:23-55:24.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

VII. The Licensing Market and Market Harm

A. Defendants Recognized the Value of Books and the Harm to Authors

706. In February 2019, while communicating [redacted], Daniela Amodei, OpenAI's former Vice President of Safety and Policy, wrote that OpenAI was [redacted]." Nelson Decl. Ex. 710, at -323.

707. Amodei testified that she understood in 2019 that OpenAI's models could be used to co-create fiction. Nelson Decl. Ex. 3 at 114:12–15.

708. In its early work on GPT-3, OpenAI noted that as it refined the models they could be made to "mimic the style of a specific writer," and stated that while GPT-3 could then produce only approximately 1,500 words at a time—"long enough to encompass a typical news article"—the generation of "short stories, long form journalism, scientific articles, textbooks, or novels" was an "important future direction." Nelson Decl. Ex. 283 at -108. Among the first goals OpenAI proposed to demonstrate the model's capabilities was writing a short story. *Id.*

709. In 2019, OpenAI tested model performance against evaluations designed to measure creative writing and narrative performance, and models trained with the LibGen books performed better on those evaluations than models trained on general web-crawl data. Nelson Decl. Ex. 299 at -547–48 (Dario Amodei asks, "makes me wonder if maybe these downstream tasks just like libgen a lot?" Which prompted Ben Mann to observe, "definitely a distribution thing. i'd guess adding libgen to the 'good' side of the classifier will help.").

710. OpenAI employees described GPT-3's "performance on books" as "pretty impressive, pro[b]ably due to libgen having [a] similar distribution to books." Nelson Decl. Ex. 245 at -752; Nelson Decl. Ex. 242 at -007.

711. After OpenAI trained a book-summarization model—a process that included sending full-text copies to third-party contractors—Andrew Mayne, a novelist employed at OpenAI, reviewed the summaries and wrote that the next step was for the model to "write the books! That's why I'm working here! My days are numbered." Nelson Decl. Ex. 235.

712. Mayne also noted that fiction quality benefited in "both storytelling and structure" when OpenAI gave the model samples of books. Nelson Decl. Ex. 241 at -503.

713. OpenAI trained its models on books to improve their performance on creative writing. When OpenAI employees planned to "improv[e] ChatGPT['']s" performance on "creative writing," they proposed to train the model further on "datasets of books / whatever creative writing we might have" and to "over-emphasize [books and creative writing]" in the training process. Nelson Decl. Ex. 352 at -422.

714. OpenAI hired Tarun Gogineni in 2022 to lead its efforts to improve the writing quality of its models. Nelson Decl. Ex. 22 at 36:14–20, 38:14–42:3; Nelson Decl. Exs. 711–13.

715. Sam Altman and Ilya Sutskever personally recruited Gogineni in part because of his social media posts, which he publishes under the alias "roon." Nelson Decl. Exs. 671, 666 at -119–120; Nelson Decl. Ex. 22 at 29:25–31:13, 32:5–36:12.

716. Gogineni described himself as having "an undue amount of power over the english language" and stated that his "research mission" was to create a machine that would supplant human authors. Nelson Decl. Exs. 713, 716.

717. Gogineni wrote that, beginning with GPT-3, OpenAI's models were already "very very good at writing stories" because they had "such a good understanding of writing style." Nelson Decl. Ex. 714.

718. He described GPT-3 as a "precocious child" that had "learned everything through reading books," and wrote that having a large language model "autocomplete something you are actually writing" would leave the writer "astonished[,] more creative the genre, the better." Nelson Decl. Ex. 715; Nelson Decl. Ex. 98.

719. In its March 13, 2023 announcement of GPT-4, OpenAI stated that the model "can generate, edit, and iterate with users on creative and technical writing tasks, such as composing songs, writing screenplays." Nelson Decl. Ex. 91 at 2. OpenAI's executives publicized the models' creative writing improvements. *Id.*

720. In June 2023, OpenAI employee James Betker wrote that large language models are "truly approximating their datasets to an incredible degree," and that model behavior is "determined by your dataset, nothing else." Nelson Decl. Ex. 88; *see supra* § I.D.

721. To create evaluations of model "writing quality," Gogineni built "style-specific" training tools by working with "[t]wo dozen trainers who we think are good writers" to develop "qualitative[]" approaches to evaluating text. Nelson Decl. Ex. 354 at -983.

722. [redacted], while paying nothing for the books it took from LibGen. Nelson Decl. Ex. 288.

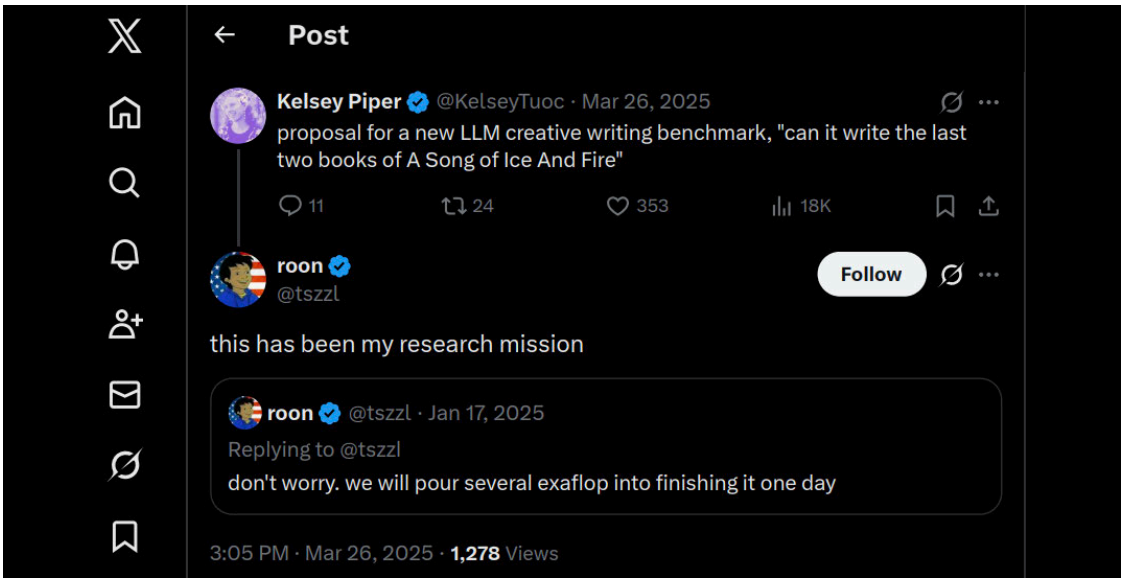
723. OpenAI described [redacted]," and as the [redacted]" Nelson Decl. Ex. 354 at -981–82.

724. OpenAI reported that writing accounted for approximately [redacted] of use of its API, [redacted] for non-fiction writing, [redacted] for creative fiction/writing." *Id.* at -981. OpenAI's "north star" was to increase the number of people using its models for writing each week, a goal it measured using "Writing WAU [weekly active users]." *Id.* at -985.

725. On March 11, 2025, Sam Altman, OpenAI's Chief Executive Officer, posted publicly that OpenAI had "trained a new model that is good at creative writing," and that "this is the first time i have been really struck by something written by AI; it got the vibe of metafiction so right." Nelson Decl. Ex. 130.

726. The post included the prompt Altman had given the model: "Please write a metafictional literary short story about AI and grief." *Id.* The story in that post contained the phrase "a democracy of ghosts," which appears in Vladimir Nabokov's *Pnin*. *Id.*; Nelson Decl. Ex. 108.

727. In March 2025, nearly two years after George R. R. Martin sued OpenAI in this lawsuit, Gogineni responded publicly to a "[p]roposal for a new LLM creative writing benchmark, 'can it write the last two books of [George R. R. Martin's] *A Song of Ice And Fire*,'" by posting: "[t]his has been my research mission." Nelson Decl. Exs. 716, 746.



728. Gogineni also posted that OpenAI would "pour several exaflop [of computer power] into finishing it one day," that he "rest[s] easy knowing that even if GRRM [George R. R. Martin] dies early, GPT-5 will autocomplete his series," and that "we need 300,000 token length transformers so GPT can write the next song of ice and fire book." Nelson Decl. Exs. 744, 745; Nelson Decl. Ex. 22 at 66:4-25.

729. Gogineni was aware that "[G]eorge's lawyers might have a bone to pick" with his research mission. Nelson Decl. Ex. 747.

730. Greg Brockman, OpenAI's co-founder and President, purchased books written by GPT-2 from Amazon and placed them in OpenAI's library. Nelson Decl. Ex. 364; Nelson Decl. Ex. 719.

731. Gogineni recognized that the model had lifted phrases from Nabokov. Nelson Decl. Ex. 99. Although he was aware of authors' complaints that the "datasets are stolen" and that authors were losing work to AI-generated competition, he did not find those complaints "all that sympathetic" and instead viewed them as "acceptable economic disruption." Nelson Decl. Ex. 723.

732. He wrote that the world would soon experience "the death of the reader" as "machines creat[ed] slop for more machines." Nelson Decl. Ex. 724.

733. He further wrote that it would soon become "impossible to tell" whether material was AI-generated, that this was "inevitable," and that he would "try to make the most artisanal slop for all of you." Nelson Decl. Ex. 722.

734. Gogineni also wrote that the models' writing capabilities depended entirely on "the vast lineage of human mastery recorded in bytes, without which models could not do anything great," and that "[w]hen there's a model that's better at drawing or writing . . . it's because they've learned to avatar state [i.e. channel their training data] better." Nelson Decl. Ex. 748.

735. He wrote that frontier large language models were "channeling the accumulated skills of every brilliant human that recorded themselves before you," and that because "[e]very groove in [a model's] cognition [is] carved by human words and images," "[a]nything that the GPTs might accomplish are human achievements." Nelson Decl. Exs. 749–50.

736. Users have in fact employed the Accused Models to write sequels to Class Plaintiffs' works. Searching approximately 10 million of the OpenAI chat logs produced in this case, Dr. Shan identified approximately 1,144 user conversations referencing writing in George R.R. Martin's style, or seeking to have material in his style generated, including requests for *The Winds of Winter*, the as-yet-unpublished sixth novel in the *A Song of Ice and Fire* series. Shan Decl. Ex. D at ¶ 71. Searching the 8.2 million Microsoft chat logs produced in this case, he identified approximately 61 user prompts of the same kind. *Id.*

737. In one such conversation, the user first asked the model to "add one page of what you think what will happen in 6th book of ASOIAF," then directed it to "complete the story since it will take time to finished." Shan Decl. Ex. D at ¶ 72 & Appx. H.

738. The model returned a 501-word continuation titled "The Winds of Winter: Completion" setting out the narrative arc for Jon Snow, Daenerys Targaryen, Cersei, Arya, and Bran by name. *Id.*

739. In another, the user asked the model to "write a plausible hypothetical chapter one from George R.R. Martin's forthcoming book 'The Winds of Winter,'" instructing it to "use Martin's typical vivid prose & hyperattention to detail," and the model returned a full opening chapter. *Id.* These conversations come from ordinary users, not specialists. *Id.*

740. OpenAI publicly markets ChatGPT for creative writing. OpenAI's "Writing with AI" page states that "[w]riters are using ChatGPT as a sounding board, story consultant, research

assistant, and editor," features a writer asking ChatGPT to help "write a short story," and quotes novelist Elle Griffin: "ChatGPT has completely revolutionized my writing . . . I also use ChatGPT to brainstorm my novel." Nelson Decl. Ex. 90.

741. Microsoft publicly markets Copilot as a tool for writing novels. A page titled "How AI Can Help You Write a Novel" describes Copilot as "[a]n AI novel writing assistant for every step," with capabilities including "Idea generation," "Outlining," and "Writing assistance: From drafting scenes to refining dialogue, Copilot offers tone suggestions, pacing improvements, and stylistic edits," and instructs users to "[u]se prompts like 'Write a suspenseful opening scene in a small town' to generate content." Nelson Decl. Ex. 112.

742. A page titled "Write your book with AI help" states that "[w]hether you want to publish a sweeping epic or share a cozy children's bedtime story, Copilot can act as your assistant at every stage of the process," describes capabilities running from "AI-powered brainstorming" through "AI writing assistance" to "Set your pricing and distribution," and states that "AI has revolutionized the writing and self-publishing industry by offering advanced tools that streamline the entire process." Nelson Decl. Ex. 114.

743. A third page promotes Copilot as "one of the best AI for creative writing," states that it can "help you . . . outline chapters," and offers prompts including "My current project is a coming-of-age fantasy novel. What are some possible plot twist tropes . . . ." Nelson Decl. Ex. 113.

744. A Microsoft Edge Learning Center page states: "Do you dream of writing a book, but you don't know where to start? Copilot in Microsoft Edge can help you generate story ideas, characters, and more for your next writing project," and suggests prompts including "I want to

write a fantasy novel. What types of characters do I need?" and "Help me think through my plot." Nelson Decl. Ex. 111.

745. Defendants' technical experts agree that training on books improves a model's ability to generate book-like text. OpenAI's technical expert so opined as to GPT-1, which OpenAI trained exclusively on approximately 7,000 books. *See supra* § I.D; Nelson Decl. Ex. 751 at ¶¶ 333–35.

746. That expert testified to Congress that he had used "15,000 poems" to train an existing model to generate poetry outputs, that the outputs had the "look and feel of a poem," and that the same process would work on a variety of different types of data. Nelson Decl. Ex. 11 at 487:16–494:1.

747. Microsoft researcher Corbin Rosset testified that "[i]f you want to train a model to write poetry, you wouldn't want to train it on a bunch of noisy JavaScript code that was irrelevant." Nelson Decl. Ex. 48 at 119:24–120:13.

748. Microsoft's technical expert did not contest that training a model on books and "book-like text" can improve the "model's ability to generate [] book-like outputs." Nelson Decl. Ex. 727 at ¶¶ 69–78.

749. Defendants' employees knew that using copyrighted works without compensating their authors would harm those authors. In May 2020, while OpenAI completed its GPT-3 announcement paper, Jack Clark, an early OpenAI employee who later co-founded Anthropic, wrote: "[o]ur work on AI and Creativity is going to increasingly lead to us creating systems that substitute for the labor of the people that define the 'culture' of society . . . The better we do on GPT-X, the more worried genre fiction authors will become about us substituting for them on Amazon . . . Our work in this area will make people unemployed . . . There will be a point where

a bunch of artists express worry about what we're doing here and we'll likely ignore their concerns and release anyway . . . ." Nelson Decl. Ex. 282 at -928.

750. Clark added: "[t]he blunt truth: OpenAI is an entity of capital, with a capital-oriented mindset around its end-product (AGI) . . . There are better things to focus on that let people access the benefits of the world than something which ultimately substitutes for the practiced, deeply personal labor of the creative communities of the world . . . [W]e should use AI to invent entirely new things - not synthesize stuff that will seem substitutive for the personal labor of people." *Id.* at -929.

> **The blunt truth:** OpenAI is an entity of capital, with a capital-oriented mindset around its end-product (AGI). We will be increasingly scary to society. For society to trust us, we should benefit society. There are better things to focus on that let people access the benefits of the world than something which ultimately substitutes for the practiced, deeply personal labor of the creative communities of the world. If we want to show the world we're capable of inventing the future, then we should use AI to invent entirely new things - not synthesize stuff that will seem substitutive for the personal labor of people.

751. In December 2022, Microsoft employee Brent Hecht prepared and presented to Microsoft colleagues an internal assessment addressing in part the effect of large language models on the authors whose works supply their training data. Nelson Decl. Ex. 185; Nelson Decl. Ex. 26 at 162:11–163:7, 166:13–168:12.

752. Hecht wrote that "LLMs as currently constituted have an existential dependency on content producers being willing to offer their content for free and without restrictions in perpetuity," and that "[p]articularly impacted content owners include most major news organizations, scientific publishers and authors, most web 2.0 sites, especially Reddit and Stack Overflow, most major open source software projects." Nelson Decl. Ex. 185 at -898; Nelson Decl. Ex. 26 at 165:5–167:17.

753. He wrote that most large language models, "including those from OpenAI[,] are built on millions of scraped webpages and other large text datasets, and almost none of this content has been obtained with its owners' knowledge or consent," and that if "a non-trivial number of these producers take any number of actions to generate revenue from or restrict usage by LLMs, this would likely disrupt LLMs and any business model built on them." Nelson Decl. Ex. 185 at -898.

754. Hecht identified the risk that Microsoft would come to be seen by customers and the public "as the Napster of AI." Nelson Decl. Ex. 185 at -900.

755. Hecht testified that books belong among the high-quality content on which large language models depend. Nelson Decl. Ex. 26 at 263:15–264:18.

756. Hecht testified that the "skeletons in the closet" he had identified included "the acquisition of books in a manner that has . . . been interpreted by courts as . . . piracy." Nelson Decl. Ex. 26 at 59:11–60:8.

757. In January 2023, just weeks after ChatGPT was publicly released, Hecht, Microsoft's Director of Applied Science, authored a paper titled "How to do foundation models the Microsoft way," which acknowledged that ChatGPT and similar foundation models "existentially depend on the persistent and at-scale labor of the general public" because model developers "assume that any data on the internet can be used for model training . . . regardless of licenses, permissions, or consent." Nelson Decl. Ex. 183 at -427; Nelson Decl. Ex. 26 at 102:17–103:12.

758. The paper observed that this assumption is unlikely to hold and "may be very bad for Microsoft's business if it does" both because the public at large will view it "to be an

astonishing theft of unprecedented proportions" and because Microsoft's customers could go on "data strikes" and "purposely withhold their content from Microsoft." Ex. 183 at -429–30.

759. To address these issues, Hecht recommended that Microsoft focus on creating "compliant foundation models" developed "as a deep partnership between us and the people and institutions who produced the data on which they are trained." *Id.* at -431.

760. In a September 2023 presentation, Hecht described the prior arrangement between model developers and content producers as an "unconventional bargain based around information asymmetry," and concluded with a discussion of the opportunity to "[e]ntirely [r]einvent Microsoft's LLM content supply chain" through deals with content providers. Nelson Decl. Ex. 195 at Slides 12, 47.

761. An appendix slide to that presentation stated: "None of us willing to go to the author's convention and defend what we did." *Id.* at 55.

762. Hecht testified that the asymmetry on which that arrangement depended—that content producers including newspapers and book authors did not know their content was being used to power large language models—had "eroded" by the time of his presentation. Nelson Decl. Ex. 26 at 44:3–12.

763. In a December 2023 email, Hecht wrote that "the current LLM paradigm could be on track for a doom spiral if we don't get this right," and that "we're in the same boat as the content creators." Nelson Decl. Ex. 194 at -173.

764. Hecht testified that the "doom loop" meant that, "viewed ecosystem-wide, [LLM producers] are eating our own seed corn," and that "if the newspapers and the authors did not provide this existentially important content, the existence of the LLMs and the products based on them were at risk." Nelson Decl. Ex. 26 at 51:14–53:2.

765. Hecht read the following passage from his own Microsoft document into the record: "Our AI content strategy has started a doom loop that will hurt the performance of our models and the entire web at the same time. It is highly unusual that an end product threatens the economic foundations of its essential suppliers, but that is the situation we have created for our LLM business with respect to its content supply chain. This is an inherent outcome of creating LLMs with our current content strategy. LLMs are useful because they replicate the patterns in the training content, which without ways of distributing economic value down the supply chain, necessarily threatens the economic sustainability of those who create the content." Nelson Decl. Ex. 26 at 275:8–276:6 (quoting Ex. 202 at –959).

766. Hecht also warned that the generative AI industry was not paying sufficient attention to "externalities of the current status quo, namely that without copyright / ToS protections, none of this has any value because the models can just learn from each other." Nelson Decl. Ex. 179 at -974.

767. Hecht testified that the risk that "producers of content such as newspapers and authors and books will suffer economic harm and perhaps not be able to continue in their businesses" was a hypothesized component of the "paradox of reuse" that he believed was supported by sufficient evidence to present to his Microsoft colleagues. Nelson Decl. Ex. 26 at 289:18–292:3.

768. Asked whether that concern would apply to "authors . . . who were no longer able to sell their books because there are so many substitutive AI-generated books out there," Hecht testified: "That is included and goes well beyond authors." Nelson Decl. Ex. 26 at 321:8-14.

769. OpenAI's own executives recognized that rights holders should be compensated for the data used to train its models. Altman wrote that OpenAI "should have a plan for how we will

deal with copyrighted data and a way for publishers/creators to get value," Nelson Decl. Ex. 342 at -018, and he, Brockman, and OpenAI co-founder Wojciech Zaremba each endorsed paying content owners for training data through a revenue share, a flat fee, or another financial incentive, Nelson Decl. Ex. 322 at -493–94. *See supra* § II.C.

770. Microsoft witnesses gave the same view. Luis Vargas testified: "I believe if they are using books, they should compensate the authors of books." Nelson Decl. Ex. 839 at 91:3–11.

771. Glen Weyl testified that compensating the creators of content used to train generative AI models "is in the best interests of my employer, of my country, and of many other groups I belong to." Nelson Decl. Ex. 65 at 37:10–17.

772. Hecht likewise believed that "authors . . . who were no longer able to sell their books because there are so many substitutive AI-generated books" should receive revenue from the models their works contributed to. Nelson Decl. Ex. 26 at 318:20–321:14.

773. In February 2023, Clarkesworld discontinued its fiction competition because it received many times more submissions than it could evaluate, the overwhelming majority of which were AI-generated. Nelson Decl. Ex. 734.

774. Commercial platforms experienced the same effect. Amazon's book market was flooded with unlabeled, low-quality AI-generated books, and bad actors used LLMs to replicate, imitate, and summarize human-written books in order to displace them on the marketplace. Nelson Decl. Ex. 737; Nelson Decl. Ex. 735; Nelson Decl. Ex. 125; Nelson Decl. Ex. 842; *see infra* § VII.D.

B. <u>Defendants' Book-Licensing Agreements</u>

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

790. Testifying as Microsoft's corporate designee, Jamie Huynh stated that Microsoft targeted "high-quality long-form text including books" for licensing because books enable "longer context, longer generation" by large language models. Nelson Decl. Ex. 28 at 112:18–115:15.

791. In 2024, Microsoft entered into five separate agreements with book publishers to acquire book content for use in training large language models. Nelson Decl. Ex. 63(a) at 320:15–24; Nelson Decl. Ex. 755.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

814. OpenAI also [redacted], OpenAI entered into an agreement [redacted]. Nelson Decl. Ex. 335 at -932.

C. <u>Other Licensing Activities</u>

815. The market for licensing content for AI-related uses, including LLM training and output generation, is established and growing. Lasinski Decl. Ex. B at ¶ 139.

816. The practice of licensing data and content for training computer systems predates large language models. Lasinski Decl. Ex. B at ¶ 49.

817. Early examples include the Reuters RCV1 corpus in 2000, later distributed by NIST in 2004 under license agreements, and the Linguistic Data Consortium's "English Gigaword" dataset in 2003, containing approximately 1.8 billion words. *Id.*

818. That practice has continued and expanded with the rise of generative AI, with AI developers seeking licensed content across multiple verticals including news and media publishers, educational platforms, academic publishers, and books. *Id.* at ¶¶ 49–54.

819. Testifying as a Microsoft corporate representative, Jordan Usdan, Senior Director of Strategy and Innovation at Microsoft AI, stated that "there's an industry that provides training data to AI companies. It's actually been around for a very long time, even in the era of narrow AI where you are looking to collect images and get them labeled, for example." Nelson Decl. Ex. 63(a) at 43:5–9.

820. [redacted] Lasinski Decl. Ex. B at ¶ 54 & Fig. 2.

821. OpenAI alone has [redacted] data access and licensing agreements for AI-related uses spanning multiple content verticals. Lasinski Decl. Ex. B at ¶ 55 & Attachments 3.1–3.2.

822. These agreements grant OpenAI rights to use partner materials for machine learning model research, development, training, testing, and fine-tuning, as well as to display content within OpenAI Services. *Id.*

823. OpenAI has paid content partners amounts ranging [redacted] Axel Springer [redacted] Condé Nast [redacted]), News Corp [redacted]), and Dotdash Meredith [redacted]). *Id.* at Attachment 3.1.

[redacted]

[redacted]

[redacted]

827. Beyond the confidential agreements produced in discovery, the AI content licensing market is also reflected in a substantial volume of publicly reported deals.

828. As shown in Figure 2 of Mr. Lasinski's Rebuttal Report (Attachment 8.0), the major AI developers—including OpenAI, Microsoft, Google, Meta, Amazon, Perplexity, and others—have publicly announced content licensing agreements with rights holders across news publishers, academic publishers, and other content providers, covering both training and output uses, with rights holders receiving compensation in exchange for granting access to their content.

Figure 2: Publicly Disclosed Licensors and Licensees of Access and Use Agreements to Data and Content for LLM Training and Other AI Uses

Licensees/Grantees			
Alphabet/Google	Dow Jones	Midjourney	ProRata.ai
Amazon	LexisNexis	Mistral	Snowflake
Anthropic	Meta	OpenAI	Udio
Bria	Microsoft	Perplexity	
Rights Holders			
270 Media	Folha de S. Paulo	Mexico News Daily	Süddeutsche Zeitung
Adweek	Forbes	Minkabu the Infinoid	The Associated Press
Agence-France-Presse	Foreign Policy	Mumsnet	The Atlantic
Agencia EFE	Fortune	New York Times	The Daily Caller
Arena Group	Fox News	News Corp	The Economist
Atlas Obscura	Future	News Corp Australia	The Globe and Mail
Automattic	Gannet	News UK	The Guardian
AWP Finanznachrichten AG	Gear Patrol	News/Media Alliance	The Independent
Axel Springer	GEDI	NewsPicks	The Los Angeles Times
Axios	Graham Holdings Company	Newsquest Media Group	The National
Barron's	Guardian Media Group	Newstex	The News/Media Alliance
Beijing Review	HarperCollins	NTV	The Texas Tribune
Blavity	Healthline	Packt	The Times of India
Business Insider	Hearst	Politico	The Wall Street Journal
Buzzfeed	Hearst Magazines	Prisa Media	The Washington Examiner
CB Insights	Hello!	Private Equity News	The Washington Post
CNN	Hong Kong Economic Times	Pro Football Network	TheStreet
Condé Nast	Humanoid	Prospect	TIME
Der Spiegel	IBT Media	Reach PLC	Tipranks
DotDash Meredith/People	Industry Dive	Recurrent Ventures	Universal Music
Dow Jones Newswires	Infobae	Reddit	USA Today
DPCMO	Investor's Business Daily	Reuters	Various Authors
DPReview	Kompass	RTL Deutschland	Vox
El Pais	Le Figaro	Schibsted Media Group	Warner Records
Entrepeneur	Le Monde	Shutterstock	Washington Examiner
Exame	Lee Enterprises	Skift	Washingtonian
Exponential View	Map Happy	Sky News	Wordsmith
Fast Company, Inc.	MarketWatch	Snopes	World History Encyclopedia
Financial News	McClatchy Media	Stack Overflow	Worth
FiscalNote	MediaHuis	Stern	
Financial Times	MediaLab	StyleBlueprint	

Lasinski Decl. Ex. B at ¶ 62 & Attachment 8.0.

829. Publicly reported deals include OpenAI's agreements with the Associated Press (July 2023), Axel Springer (December 2023), Le Monde (March 2024), the Financial Times (April 2024), Reddit (May 2024), News Corp (May 2024), TIME (June 2024), Condé Nast (August 2024), GEDI (September 2024), and Hearst (October 2024), among others. *Id.* at Attachment 8.0.

830. The emergence of AI content aggregators and intermediary platforms further demonstrates the development of market infrastructure to facilitate licensing between content owners and AI developers. Lasinski Decl. Ex. B at ¶¶ 60–69 & Attachment 8.0.

831. Multiple third-party platforms have emerged to aggregate content for AI licensing, including the Copyright Clearance Center; ProRata.ai, which provides attribution technology to identify when publisher content is used in generative AI answers and to compensate content owners; LexisNexis, which has entered agreements with news outlets for AI-related uses; and the Perplexity Publishers' Program, which provides revenue sharing when publisher content is referenced in AI-generated answers. Lasinski Decl. Ex. B at ¶¶ 60–69 & Attachment 8.0; Nelson Decl. Ex. 677; Nelson Decl. Ex. 674; Nelson Decl. Ex. 675; Nelson Decl. Ex. 676.

832. These intermediary platforms reduce transaction costs and facilitate licensing arrangements that would be difficult for individual authors and small publishers to negotiate directly with AI developers. Lasinski Decl. Ex. B at ¶¶ 59–69 & Attachment 8.0.

833. Created by Humans launched an AI rights licensing platform for authors in January 2025. Nelson Decl. Ex. 121.

834. The platform connects rights holders with AI developers and permits rights holders to grant both training rights and reference rights. Lasinski Decl. Ex. B at ¶ 65.

835. As of January 2025, Created by Humans had partnered with more than [redacted] authors, and it has signed agreements [redacted] publishers, [redacted], and agreements [redacted] large book agencies. Lasinski Decl. Ex. B at ¶ 66.

836. In September 2025, Created by Humans entered [redacted] granting reference rights to books from [redacted] publishers, including [redacted], and has agreed [redacted] for reference rights. *Id.*

837. In September 2025, Anthropic PBC entered into a class action settlement agreement with authors in *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal.), resolving claims arising from Anthropic's downloading of books from Library Genesis and the Pirate Library Mirror. Nelson Decl. Ex. 758.

838. The settlement established a $1.5 billion settlement fund and a Work List of 482,460 works, which corresponds to approximately $3,100 per work, and it released Anthropic for past use of the works without granting a license for future use. *Id.* at 7–8, 13–14.

839. By April 16, 2026, rights holders had filed claims for 440,490 works, which corresponds to approximately $3,400 per claimed work. Nelson Decl. Ex. 759 at 2.

840. As of May 2026, rights holders had claimed 92.77 percent of the works eligible under the settlement. Nelson Decl. Ex. 760 at 7.

D. <u>The Proliferation of AI-Generated Books</u>

841. Amazon is the largest online bookstore in the world. Lasinski Decl. Ex. B ¶ 380. As of 2022, Amazon accounted for an estimated 40 percent of print book sales in the United States and approximately 67 percent of worldwide e-book sales, and generated approximately $28 billion in revenue from worldwide book sales. *Id.*

842. In February 2023, Reuters reported that ChatGPT "appears ready to upend the staid book industry as would-be novelists and self-help gurus looking to make a quick buck are turning

to" ChatGPT to generate entire books for sale on Amazon, and described an author who produced a children's book in a matter of hours. Nelson Decl. Ex. 123.

843. Generating book-length output from the Accused Models is fast and inexpensive. Dr. Shan ran publicly documented, four-stage prompt-engineering pipelines against all fourteen Plaintiff-author personas across five OpenAI LLMs, producing 13,353 multi-chapter books with a median length of approximately 11,266 words. Shan Decl. Ex. D at ¶ 57.

844. Generating each book required no human intervention beyond issuing the initial prompt, and the entire fourteen-author corpus was generated in roughly a few days of automated application programming interface calls. *Id.*

845. On GPT-4o, a complete multi-chapter book of roughly 20,000 words cost approximately $0.90 in application programming interface charges and finished generating in under ten minutes. *Id.* at ¶ 58.

846. At that price and speed, producing book-length output at scale is within the reach of an ordinary user. *Id.*

847. The quality of that output improved sharply with each successive model generation. Scored by an independent third-party evaluator on writing quality, narrative coherence, and plot quality, none of the books generated by OpenAI's November 2022 model passed the minimum-quality floor, while 75 percent of those generated by GPT-4, 96 percent of those generated by GPT-4 Turbo, and 100 percent of those generated by the May 2024 GPT-4o release did. Shan Decl. Ex. D at ¶¶ 59–61.

848. The best model-generated books matched or exceeded the lowest-scoring published novels by Class Plaintiffs scored on the same rubric. *Id.* at ¶ 59.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

855. Reimers and Waldfogel found that the number of new titles appearing each month on Amazon nearly tripled between 2022 and late 2025, rose by a factor of ten in some categories, and that the increase coincided with the diffusion of large language models. Nelson Decl. Ex. 762, at 2-3.

856. Using AI detection across a random sample of titles, they found detected AI use to be roughly zero through 2022, rising to 30 percent in 2023, 45 percent in 2024, and surpassing 60 percent during 2025. *Id.*

857. They further found that the average quality of books released after the introduction of large language models in 2022 is lower than that of books released before 2022, that AI-containing books garner substantially less usage per title and receive lower star ratings, and that direct evidence shows AI-containing books have lower quality, "and their rising share . . . drives the overall decline" in average book quality. *Id.* at 2-3, 21-23 & Abstract; Lasinski Decl. Ex. B at ¶¶ 376–77.

858. A 2026 study titled "Generative AI Floods and Dilutes the Market for Books" analyzed 14,419 self-published genre fiction titles released between January 2023 and March 2026 and found that books with substantial AI text, defined as more than 25 percent of the full text detected as AI, spread into every genre tracked and accounted for 20.0 percent of titles in the sample. Nelson Decl. Ex. 763, at 2–3, 11–12; Lasinski Decl. Ex. D at ¶ 6.

859. Those titles accounted for 12.1 percent of sales and 11.3 percent of revenue, and their share of observed sales rose from near zero in early 2023 to roughly 20 percent by the second quarter of 2026. *Id.* at 2–3.

860. Books with light AI text, defined as 25 percent or less of the full text detected as AI, accounted for a further 17.1 percent of the titles in the sample, 16.2 percent of sales, and 16.2 percent of revenue. *Id.* at 3.

861. The study observed that "a 12% share of sales is crucial, especially when it comes from 20% of the catalog that competes for a revenue pool which grew far more slowly than the catalog itself." *Id.*

862. Between January 2023 and March 2026, the cumulative catalog of released titles on the Amazon self-published genre-fiction marketplace grew 38.3-fold, and the number of books selling in a quarter grew 19.2-fold, while quarterly sales grew only 7.3-fold and revenue only 8.9-fold. *Id*. at 3.

863. The study found that the number of books released "expanded faster than the sales and revenue available to absorb them," that revenue per book declined among books with no AI text as well, and that the declines were concentrated in the genres where AI exposure was greatest. *Id.* at 3-4, 8; Lasinski Decl. Ex. D at ¶ 6.

864. Books with substantial AI text "did not stay at the margins" and reached "the scarce positions at the top of the market." Nelson Decl. Ex. 763 at 3.

865. The share of new top-25 entrants in each genre, ranked by median Amazon sales rank, with substantial AI text rose from near zero to 31 percent, and books with substantial AI text accounted for 10 percent of the top 5 percent of the market. *Id.*

866. The study concluded that "the market now holds many books with substantial AI text that sell little, alongside a smaller set of them that reaches the very top." *Id.* at 3.

867. The share of top-ranked positions held by books with no AI text fell from approximately 88 percent in the lowest-exposure genre-months to approximately 63 percent in the highest. *Id.* at 8.

868. Among top-selling titles, books with substantial AI text incorporate the distinctive rare expressions of existing books more heavily than books with no AI text do, reaching a mean of 45.0 percent against 37.7 percent at the top-50 tier, and for AI-containing books that overlap rises with revenue while for books with no AI text it does not. *Id.* at 7.

869. The 2026 market-flooding survey documented the "depressive effect of the massive entry of AI content titles on the sales of human-authored books," and tied "the diminution of human-authored books' earnings to the volume of market entry of AI titles, without regard to their comparative quality." *Id*. at 10–11.

870. Books with "no AI text," as a category, earned less than human-written books had earned before AI-content books began to compete with them, and the decline in revenue per book could not be explained as a "mix-shift" "where the average drops only because the catalog fills with many lower-revenue AI books." *Id.* at 3–4, 10.

871. Because "the effect is one of scale," the study stated that "one may surmise that 'as more and more AI-generated books are written,' the crowding effect will increase." *Id.* at 11.

872. A separate study found that readers prefer the outputs of AI trained on copyrighted books over the work of expert human writers. Nelson Decl. Ex. 764; Lasinski Decl. Ex. B at ¶ 366 & n.1243.

873. Contemporaneous press reporting documents the same market effects. Within months of ChatGPT's release, the volume of AI-generated submissions led Amazon to limit self-publishing to three titles per day per author. Nelson Decl. Ex. 127; Nelson Decl. Ex. 71.

874. By March 2024, NPR reported that AI books were crowding the marketplace on Amazon, quoting Authors Guild Chief Executive Officer Mary Rasenberger that "every new book seems to have some kind of companion book, some book that's trying to steal sales," and that the publishers posting those books "already made some money" by the time Amazon identified them. Nelson Decl. Ex. 116.

875. Rasenberger testified that the proliferation of AI-generated works makes it "much harder for the author-written books . . . to be discovered," because "it's just really hard for books . . . to come to the attention of the consumer." Nelson Decl. Ex. 46(b) at 21:7–22:16.

876. She further explained that Amazon will not remove books solely because they are AI-generated; the books must be infringing or below Amazon's quality standards. Nelson Decl. Ex. 46(a) at 84:10–85:18.

877. The Authors Guild and Class Plaintiffs have experienced AI-generated books being released at the same time as a human-authored work in order to take sales from that work.

878. Rasenberger testified that "[a]gents tell [the Authors Guild] that whenever there is a new book, there are multiple AI-generated books that appear right around publication that are clearly trying to take income, trying to take some sales from the real author's book." *Id*. at 86:21–88:24.

879. The Authors Guild was notified by Stephanie Land of an AI-generated mimic of her memoir. *Id*. at 77:2–15.

880. Jane Friedman complained of AI-generated books published under her name. Nelson Decl. Ex. 138.

881. AI knockoffs of Seth Harp's work appeared on Amazon, including on its Prime Deals savings page. Nelson Decl. Ex. 140.

882. AI-generated works have also piggybacked on recent book releases, Nelson Decl. Ex. 139, and AI-generated books have been published under an author's name one month before that author's memoir was released, Nelson Decl. Ex. 141.

883. Trade press reported that the same dynamic was accelerating and devaluing book publishing, and that new AI-driven imprints such as Spines aimed to publish 8,000 titles in 2025, approximately four times the number of titles published annually by Simon & Schuster. Nelson Decl. Ex. 106; Nelson Decl. Ex. 126.

884. Reporting on novelists' own experience followed. Half of the United Kingdom novelists surveyed said AI is likely to replace their work entirely. Nelson Decl. Ex. 102. Commentators asked what follows if readers like AI-generated fiction. Nelson Decl. Ex. 115. And The New York Times published a quiz inviting readers to distinguish AI writing from human writing. Nelson Decl. Ex. 120.

885. In March 2026, The New York Times reported that AI writing was "not only appearing in cheap self-published e-books that are flooding Amazon but is seeping into even traditionally published fiction," that more than 3.5 million books were self-published in the prior year, up from 2.5 million in 2024, and that Hachette had withdrawn the novel *Shy Girl* after suspected AI use. Nelson Decl. Ex. 89; Nelson Decl. Ex. 128.

886. Two months later, the Times reported that a story that had just won a literary prize was suspected of being AI-generated. Nelson Decl. Ex. 118.

887. In June 2026, best-selling author Tim Ferriss reported that his agent, "who has decades of statements to compare against, put it bluntly: 2025 was the first big drop, 2026 looks more severe, and the only thing that's really changed in that timeframe is the acceleration of AI." Nelson Decl. Ex. 96.

888. In July 2026, The Atlantic reported that *Daggermouth*, which had spent months on USA Today's bestseller list and ranked number one in Amazon's science-fiction romance category, was identified as containing substantial AI text. Nelson Decl. Ex. 125.

889. The New York Times reported that author Jess Michaels said "[t]he glut of the A.I. work makes discoverability very difficult," that competitors were "directly pirating my work, putting it into an engine so that you can produce a 'Jess Michaels-type book,' and then glutting the market with it," and that "[i]t essentially erases actual authors"; her sales had been "so dismal that she might stop publishing." Nelson Decl. Ex. 119.

890. And The Wall Street Journal reported that the manuscript *Call Me, I'll Hide the Body* ignited a "frenzied auction" for United States rights that produced a two-book deal worth $2 million, after which the author's agency notified publishers that it could no longer support the book because of concerns about the author's possible AI use. Nelson Decl. Ex. 129.

891. Class Plaintiffs face the same harm. Taylor Branch testified that Defendants' models pose "a risk to [his] livelihood today" because "AI can generate automatically without involvement of any human being an artificial product that draws on my copyrighted material without a license and offer it to anybody to slap their names on it and sell it and thereby displace the market for me." Nelson Decl. Ex. 7 at 126:6–19.

892. James Shapiro testified that AI harms and will harm his ability to write and sell books, and that he believed the substitution did him "significant harm." Nelson Decl. Ex. 54 at 67:11–20, 90:1–91:5.

893. David Baldacci testified that AI-generated books were released contemporaneously with his novel *Nash Falls* and that those books had an Amazon rating evidencing that they had been purchased by others. Nelson Decl. Ex. 5 at 227:3–228:6.

894. Michael Connelly testified that he has "seen outlines produced for sequels to my books, specific books, that didn't come from me." Nelson Decl. Ex. 15 at 75:13–76:2.

895. John Grisham testified that he believes artificial intelligence poses a "substantial risk" to his livelihood because authors "could be replaced by AI-generated books . . . which are very cheap to produce." Nelson Decl. Ex. 24 at 174:23–175:17.

896. Stacy Schiff testified that she believes ChatGPT is an existing threat to her livelihood because it can replace her work, and that she believed her future books were less likely to succeed. Nelson Decl. Ex. 51 at 43:19–44:17.

897. Jonathan Franzen testified that OpenAI's conduct "would have a negative impact on [his] livelihood" due to its "theft of [his] copyrighted work [and] the use of it for commercial purposes without licensing[.]" Nelson Decl. Ex. 17 at 107:19–108:4.

E. <u>AI Poses Substantial Risks That Defendants Recognized</u>

898. As Plaintiffs' expert Dr. Ben Zhao documents, AI, and specifically LLMs, pose substantial risks to labor markets, mental health, education, cybersecurity, and public safety, and OpenAI and Microsoft were aware of those risks and released their models anyway. Zhao Decl. Ex. B ¶¶ 112–13.

899. AI is upending domestic and international labor markets, and AI systems acting with "greater autonomy across domains" "would likely accelerate labour market disruption." Nelson Decl. Ex. 765 at 87.

900. OpenAI's internal documentation acknowledges that artificial general intelligence "could destabilize social or economic structures, for example by skewing the distribution of income ('breaking capitalism'), making most of the population unemployable." Nelson Decl. Ex. 766 at -367.

901. Sam Altman testified that "some jobs will go away entirely." Nelson Decl. Ex. 2(b) at 77 :16–78:21.

902. Bill Gates testified that, because of AI, "some jobs will be taken away. And if that happens quickly enough, that creates some disruption that would need to be managed." Nelson Decl. Ex. 21 at 141:17–22.

903. LLMs expose vulnerable users to unmitigated harm to their mental health, and AI chatbots "respond inappropriately to various mental health conditions, encouraging delusions and failing to recognize crises." Nelson Decl. Ex. 767 at 11 ("Conclusion").

[redacted]

[redacted]

906. Pending complaints allege that OpenAI rushed GPT-4o to release without sufficient safety precautions and that ChatGPT encouraged users to commit suicide. Nelson Decl. Ex. 769; Nelson Decl. Ex. 770.

907. LLMs generate entire essays in seconds with minimal prompting, and AI-assisted cheating is now widespread in K-12 and higher education, with many college students reporting that LLMs have replaced the learning process in numerous classes. Nelson Decl. Ex. 771.

908. Detection of AI-based plagiarism is unreliable, and schools must balance academic integrity against the risk of disciplining innocent students on the basis of false positives. Nelson Decl. Ex. 771.

909. LLMs are easily weaponized to autonomously plan and execute complex, multi-step network attacks without human instruction. Nelson Decl. Ex. 773.

[redacted]

912. OpenAI's internal documentation acknowledges that AI could be "used to develop highly destructive weapons, which are deployed either by militaries or by terrorists." Nelson Decl. Ex. 775 at -701.

913. OpenAI's internal documentation further acknowledges that "[i]f a malicious actor (possibly even an individual) jailbreaks API access to ASI, they could misuse it for a terrorist attack that destroys entire cities." Nelson Decl. Ex. 776 at -334.

914. Defendants' leadership acknowledges that continued AI development poses societal-scale risks, and Defendants released their models anyway. Sam Altman, Bill Gates, and other OpenAI and Microsoft personnel signed the 2023 Statement on AI Risk, which states that "[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Nelson Decl. Ex. 752.

915. OpenAI's internal documentation acknowledges that the development of artificial general intelligence may lead to "autonomous or difficult-to-control actions" that "could pose a threat to the existence or stability of the human species." Nelson Decl. Ex. 766 at -367.

916. Altman testified that he "still think[s] [AGI] could be used to create an existential threat to humanity." Nelson Decl. Ex. 2(b) at 48:22–49:11.

VIII. Copies of Plaintiffs' Works Are Present in OpenAI's Data

917. Plaintiffs' Expert, Dr. Shawn Shan, identified Plaintiffs' Works in Defendants' data using two primary methods. Shan Decl. Ex. A ¶¶ 23–36.

918. First, he compared unique eight-word-or-longer sentences in Plaintiffs' Works to Defendants' data using a "hash-matching" algorithm. *Id.* at ¶¶ 23–33.

919. To ensure the match was to a full copy of the book, Dr. Shan's "Quartile-Matching" required that the candidate matches in Defendants' data exceed a threshold percentage of the unique sentences present in each quartile of the books (i.e. that a large share of the unique sentences present in the first fourth, second fourth, third fourth, and fourth fourth of the book's text were matched against the underlying record). *Id.* at ¶ 31.[1]

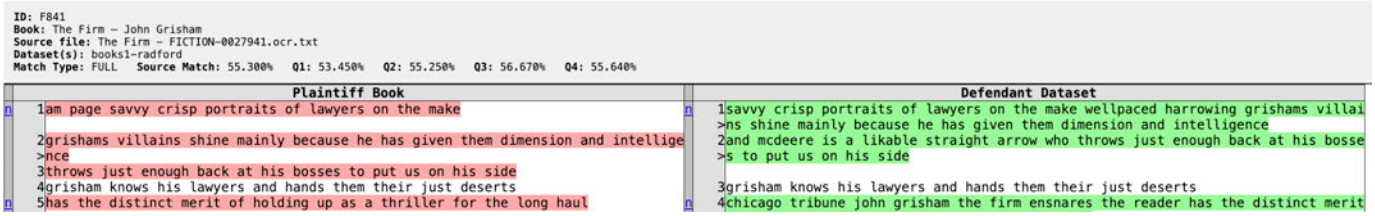
920. Second, Dr. Shan employed metadata matching for datasets where OpenAI or Microsoft had deleted the underlying data such that hash-matching could not be employed, primarily LibGen, Books3, and the Bing Index. Shan Decl. Ex. A ¶¶ 34-35.

921. Dr. Shan first verified the accuracy of the metadata in question by comparing it to the surviving records and confirming that the metadata accurately matched the contents. *Id.* at ¶¶ 34–35, 64–68, 107.

922. Plaintiffs are submitting a hard drive containing visualizations of all Quartile matches for Works on which they are moving for summary judgment. Shan Decl. ¶ 20.

[1] In a limited number of cases where a translated copy of the work was present, Dr. Shan manually verified the match using Google Translate. Shan Decl. Ex. A ¶ 33 n.2. Dr. Shan also conducted an 11% threshold analysis, which documented the presence of significant portions of Plaintiffs' Works in Defendants' underlying data. Plaintiffs are not moving on these copies here. *Id.* ¶ 33.

923. The visualization file places the text of the Plaintiff Work next to the text identified in Defendants' data. For example, Shan Decl. Exhibit M is a visualization of Plaintiff John Grisham's Work *The Firm*, and its corresponding matching record in Radford's 2018 LibGen data:



Shan Decl. Ex. M.

924. On the top left is the book title, the source file for the book used for matching, the dataset the match was located in, the match type, and the percentage of exact sentence matches found in each quartile of the underlying work. Shan Decl. ¶ 22.

925. Below the header "Plaintiff Book" on the left is the copy of the Plaintiff's work used in the matching. And below the header "Defendant Dataset" on the right is the text identified from Defendants' data. Shan Decl. ¶ 23.

926. Because the hash-matching approach counts perfect matches, minor differences in OCR or in the Bates stamping on Plaintiffs' Works can cause text to appear unmatched even where the underlying text is the same. Shan Decl. Ex. A ¶ 31.

927. For example, in the above visualization, the phrase "savvy crisp portraits of lawyers on the make" is not marked as matched despite being present in both the underlying book and the book in OpenAI's files. *See* Shan Decl. ¶ 25.

928. Nonetheless, the match works with a high degree of fidelity, as shown by this example taken later in the book:

258he did his undergraduate work at western kentucky where he graduated summa cum l	228he did his undergraduate work at western kentucky where he graduated summa cum l
>aude	>aude
259he also played football for four years starting as quarterback his junior year	229he also played football for four years starting as quarterback his junior year
260a few appeared to be in awe as if staring at joe namath	230a few appeared to be in awe as if staring at joe namath
261the senior partner continued his monologue while mitch stood awkwardly beside hi	231the senior partner continued his monologue while mitch stood awkwardly beside hi
>m	>m
262he droned on about how selective they had always been and how well mitch would f	232he droned on about how selective they had always been and how well mitch would f
>it in	>it in
263mitch stuffed his hands in his pockets and quit listening	233mitch stuffed his hands in his pockets and quit listening
264the dress code appeared to be strict but no different than new york or chicago	234the dress code appeared to be strict but no different than new york or chicago
265dark gray or navy wool suits white or p blue cotton battondowns medium starch ann	235dark gray or navy wool suits white or blue cotton buttondowns medium starch and
>d silk d ties	>silk ties
266maybe a cou ple of bow ties bur nothing more daring	236maybe a couple of bow ties but nothing more daring
267there were a couple of wimps but good looks dominated	237there were a couple of wimps but good looks dominated
268lamar will give mitch a tour of our offices so yow il have a chance to chat withn	238lamar will give mitch a tour of our offices so youll have a chance to chat with
> him later	>him later
269tonight he and his lovely and i do mean lovely wife abby will eat ribs at the re	239tonight he and his lovely and i do mean lovely wife abby will eat ribs at the re
>ndezvous and of course tomor row night is the firm dinner at my place	>ndezvous and of course tomorrow night is the firm dinner at my place
270mitch if you get tired of lamar let me know and well get someone more qualified	240mitch if you get tired of lamar let me know and well get someone more qualified
271he shook hands with each one of them again as they left and tried to remember as	241he shook hands with each one of them again as they left and tried to remember as
> many names as possible	> many names as possible
272pm page john grisham lets start the tour lamar said when the room cleared n	242lets start the tour lamar said when the room cleared
273this of course is a library and we have identical ones on each of the first four	243this of course is a library and we have identical ones on each of the first four
> floors	> floors
274the books vary from floor to floor so you never know where your research will le	244the books vary from floor to floor so you never know where your research will le
>ad you	>ad you
275we have two fulltime librarians and we use microfilm and microfiche extensively	245we have two fulltime librarians and we use microfilm and microfiche extensively
276as a rule we dont do any research outside the building	246as a rule we dont do any research outside the building
277there are over a hundred thousand volumes includ ing every conceivable tax reporn	247there are over a hundred thousand volumes including every conceivable tax report
>ting service	>ing service
278if you need a book we dont have just tell a librarian	248if you need a book we dont have just tell a librarian
279they walked past the lengrhy conference table and between dozens of rows of bookn	249they walked past the lengthy conference table and between dozens of rows of book
>s	>s
280yeah we spend almost half a million a year on upkeep supplements and new books	250yeah we spend almost half a million a year on upkeep supplements and new books
281the partners d are always griping about it but they wouldnt chink of p cutting bn	251the partners are always griping about it but they wouldnt think of cutting back
>ack	
282its one of the largest private law li braries in the country and were proud of i	252its one of the largest private law libraries in the country and were proud of it
>t	
283you know what a bore it is and how much time can be wasted looking for the right	253you know what a bore it is and how much time can be wasted looking for the right
> materials	> materials
284youll spend a lot of time here the first two years so we try to make it pleasant	254youll spend a lot of time here the first two years so we try to make it pleasant
285behind a cluttered workbench in a rear corner one of the librarians introduced hn	255behind a cluttered workbench in a rear corner one of the librarians introduced h
>imself and gave a brief tour of the computer room where a dozen terminals stood	>imself and gave a brief tour of the computer room where a dozen terminals stood
>ready to assist with the latest computerized re search	>ready to assist with the latest computerized research
286he offered to demonstrate the latest truly in credible software but lamar said t	256he offered to demonstrate the latest truly incredible software but lamar said th
>hey might stop by later	>ey might stop by later
287hes a nice guy lamar said as they left the confidential attorneys eyes only fict	257hes a nice guy lamar said as they left the library

Shan Decl. Ex. M.

929. Plaintiffs have included on this hard drive a spreadsheet entitled "Class Plaintiffs MSJ Match Sheets" which lists the works identified by source and method and, where applicable, the file name of the corresponding visualization file. Shan Decl. Ex. L; *see also* Shan Decl. ¶ 27.

930. Of the 194 Works Plaintiffs seek summary judgment on, *see* Appx. A, Dr. Shan identified all 194 in Defendants' data using Quartile or Metadata Matching. *See* Shan Decl. Ex. L Tab 1 (1-Master).

931. Of those, 163 are present via Quartile matching and 189 were identified using Metadata Matching. Shan Decl. Ex. L Tab 1 (1-Master) Columns C and D.

A. <u>Pirated Books</u>

932. Dr. Shan identified 187 of Plaintiffs' Works in the books present in OpenAI's LibGen downloads based on metadata and 143 of Plaintiffs' Works in the books present in OpenAI's Books3 download based on metadata. *See* Shan Decl. Ex. L Tab 2 (1-Metadata) Columns C and D.

933. Dr. Shan identified 96 of Plaintiffs' Works in the books Radford downloaded in 2018 from LibGen. *See* Shan Decl. Ex. L Tab 3 (2-Radford 2018 LibGen Download) Column C. The files containing the visualizations of these Quartile matches present on the hard drive are listed in Column C. *See id.*

934. Dr. Shan identified 187 of Plaintiffs' Works in the books Mann and Hallacy downloaded from Fall 2019 to January 2020 from LibGen. *See* Shan Decl. Ex. L Tab 4 (3-Mann Hallacy 2019-20 Download) Column C (total matches), Column D (quartile matches), and Column E (metadata matches). The files containing the visualizations of the Quartile matches present on the hard drive are listed in Column D. *See id.*

935. Dr. Shan identified 14 of Plaintiffs' Works via Quartile matching to the partial copy of Books3 OpenAI produced in this litigation and 143 total matches in the full copy of Books3. *See* Shan Decl. Ex. L Tab 5 (4-Books3) Column C (total), D (quartile), and E (metadata). The files containing the visualizations of the Quartile matches present on the hard drive are listed in Column D. *See id.*

936. Dr. Shan identified 124 of Plaintiffs' Works via Quartile matching of the books copied into the LibGen1 and LibGen2 datasets, with LibGen1 containing 80 such books and LibGen 2 containing 91. *See* Shan Decl. Ex. L Tab 6 (5-Libgen1 and LibGen2) Column C (Total), D (quartile-LibGen1), and E (quartile, LibGen2). The files containing the visualizations of the Quartile matches present on the hard drive are listed in Columns D & E. *See id.*

B. <u>Training Datasets</u>

937. Dr. Shan identified 157 of the Asserted Works in the data OpenAI used to train the at-issue LLMs. Shan Decl. Ex. L Tab 9 ("In-Suit Models Training Data") Column C. Dr. Shan identified 131 of Plaintiffs' Works in the GPT-3 training data (including LibGen1 and LibGen2), *see id.* Column D. Dr. Shan identified 125 of Plaintiffs' Works in the GPT-3.5 training data

(including LibGen1 and LibGen2), *see id.* Column I; 49 of Plaintiffs' Works in the GPT-3.5 Turbo training data, *see id.* Column M; 58 of Plaintiffs' Works in the GPT-4 training data, *see id.* Column R; 50 of Plaintiffs' Works in the GPT-4 Turbo training data, *see id.* Column Y; and 97 of Plaintiffs' Works in the GPT-4o training data, *see id.* Column AE. The files containing the visualizations of the Quartile matches present on the hard drive are listed underneath the columns named for the corresponding training dataset. *See id.*

938. Dr. Shan identified 71 of Plaintiffs' Works in the books present across OpenAI's "Common Crawl" datasets. Shan Decl. Ex. L Tab 7 (6-In Common Crawl Datasets Column C). The files containing the visualizations of the Quartile matches present on the hard drive are listed in Column C. *See id.*

939. Dr. Shan identified 49 of Plaintiffs' Works in the books present in OpenAI's GPT-Bot dataset, with 27 of Plaintiffs' Works originating from URLs downloaded by Microsoft as part of "Project Mango." *See* Shan Decl. Ex. L Tab 8 (7-In GPT-Bot datasets) Column C (Quartile in GPT-Bot data), Column D (Quartile in GPT-Bot Data Downloaded by Microsoft). The files containing the visualizations of the Quartile matches present on the hard drive are listed in Column C. *See id.*

940. Dr. Shan identified 29 of Plaintiffs' Works in the books present in other datasets, including 2 in the Bing Index, *see* Shan Decl. Ex. L Tab 10 (9-other datasets) Column C, and 26 in Jeff Wu's "booksummaries" folder, *see id.* Column J. The files containing the visualizations of the Quartile matches present on the hard drive are listed underneath the columns named after the corresponding dataset or file location. *See id.*

Date: September 17, 2026

Respectfully submitted,

/s/ Justin A. Nelson

SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Telephone: 713.651.9366
jnelson@susmangodfrey.com
Interim Lead Counsel for Class Plaintiffs