UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

In Re: OpenAI, Inc. Copyright Infringement Litigation This Document Relates To: No. 1:23-cv-08292-SHS-OTW No. 1:23-cv-10211-SHS-OTW No. 1:25-cv-03482-SHS-OTW	Case No. 1:25-md-3143 (SHS) (OTW) **CONTAINS INFORMATION DESIGNATED AS CONFIDENTIAL UNDER THE STIPULATED PROTECTIVE ORDER** PUBLIC VERSION – REDACTED

DEFENDANTS' RESPONSE TO CLASS PLAINTIFFS' RULE 56.1 STATEMENT

TABLE OF CONTENTS

I. The Parties, Asserted Works, and Accused Models 4
A. Class Plaintiffs and Asserted Works 4
1. David Baldacci and Columbus Rose, Ltd. 5
2. Taylor Branch 6
3. Michael Connelly and Hieronymus Inc. 7
4. Sylvia Day and Sylvia Day LLC 8
5. Jonathan Franzen 9
6. Christopher Golden and Daring Greatly Corporation 10
7. Andrew Sean Greer 11
8. John Grisham and Belfry Holdings, Inc. 12
9. David Henry Hwang 13
10. George R.R. Martin & Wo & Shade, LLC 14
11. Jodi Picoult 15
12. Stacy Schiff and Gattopardo, LLC 16
13. James Shapiro 17
14. The Authors Guild (Estate of Mignon Eberhart) 18
B. The OpenAI Entities 20
C. OpenAI's Conversion from Nonprofit to Commercial Enterprise 23
D. Generative AI and Large Language Models 27
E. The Accused Models and ChatGPT 34
F. Microsoft 41
II. OpenAI Downloaded and Stored Millions of Pirated Books from Libgen It Could Have Bought 45
A. Library Genesis 45
B. July–November 2018: OpenAI Priced Lawful Book Acquisition and Chose LibGen Instead 51
C. November 2018–February 2019: OpenAI Budgeted to Buy Books, Then Built Its First LibGen Dataset Instead 60
D. February–July 2019: OpenAI Presented LibGen to Microsoft and Closed a $1 Billion Investment 67
E. September 2019–January 2020: Mann and Hallacy Torrented the Entire LibGen Collection 74
F. 2019–June 2022: OpenAI Stored the Pirated Books in Locations Accessible to Its Employees and Put Them to Continued Use 100

1. Where OpenAI Stored the Books 100
2. Who Could Access the Books 104
3. Copies, Migration, and Retention 108
4. Continued Use of the LibGen Books 111
G. April–May 2020: OpenAI Renamed LibGen "Books1" and "Books2" 116
H. April 2021–2024: OpenAI Repeatedly Priced Lawful Alternatives and Never Replaced the LibGen Books 124
I. December 2021–Summer 2022: OpenAI Deleted LibGen 136
J. Throughout, OpenAI Knew LibGen Was "Sketchy [as F---]" 147
1. OpenAI's Contemporaneous Knowledge 147
2. Defendants' Admissions That Piracy Is Wrong and Illegal 170
III. OpenAI Also Downloaded Books from Books3, Targeted Scrapes, and Web Crawls 186
A. Books3 and The Pile 186
B. OpenAI Deleted Its Copy of The Pile in October 2024 199
C. OpenAI Also Acquired Books Through Targeted Scrapes and Third-Party Web Crawls 203
IV. OpenAI Trained Its LLMs on Datasets Containing Books 215
A. OpenAI Knew Books Were High-Quality Training Data 215
B. Defendants' Experts Acknowledge That Books Were Not Necessary 237
C. OpenAI Used LibGen to Train GPT-3 and GPT 3.5 247
D. OpenAI Also Wanted to Use Books3 to Train Its Models 252
E. OpenAI Also Trained the Accused Models on Datasets Derived from Its Scrapes and Crawls 252
F. OpenAI Created Multiple Copies of Books in Training 253
G. The Accused Models Reproduce Class Plaintiffs' Works to End Users 260
V. OpenAI and Microsoft Distributed Books to Each Other and Third Parties 279
A. OpenAI Provided Microsoft Access to GPT-3 Training Data 280
B. Microsoft Transferred Crawled Data to OpenAI (Project Mango) 302
C. OpenAI Transferred LibGen to Book-Summarization Contractors 312
VI. Microsoft's Role in OpenAI's Infringement 315
A. OpenAI and Microsoft Raced to Close Google's Lead in Compute, Talent, and Books 315
B. Microsoft Formed a Partnership with OpenAI 325
C. Microsoft Knew OpenAI Used LibGen 334

D. Microsoft Built Bespoke Supercomputing Systems for OpenAI's Model Training 345
E. Microsoft Encouraged OpenAI's Data Practices, Embedded Its Engineers at OpenAI, and Obtained OpenAI's Technology 352
F. Microsoft Could Control OpenAI's Training Data 359
G. Microsoft Benefitted Financially from OpenAI's Infringement 365
VII. The Licensing Market and Market Harm 369
A. Defendants Recognized the Value of Books and the Harm to Authors 369
B. Defendants' Book-Licensing Agreements 420
C. Other Licensing Activities 442
D. The Proliferation of AI-Generated Books 461
E. AI Poses Substantial Risks That Defendants Recognized 492
VIII. Copies of Plaintiffs' Works Are Present in OpenAI's Data 507
A. Pirated Books 515
B. Training Datasets 517
IX. Defendant OpenAI's Additional Facts Pursuant to Local Rule 56.1(b) 518
X. Defendant Microsoft's Additional Facts Pursuant to Local Rule 56.1(b) 527

Pursuant to Local Civil Rule 56.1, Defendants OpenAI and Microsoft respectfully submit this Response to Plaintiffs' Corrected Statement of Undisputed Material Facts in Support of their Motion for Partial Summary Judgment. ECF No. 1988 (the "Statement").

Plaintiffs' Statement fails to comply with Local Civil Rule 56.1(a). That rule requires that a motion for summary judgment "be accompanied by a separate, **short, and concise statement**, in numbered paragraphs, of the **material facts** as to which the moving party contends there is no genuine issue to be tried." Local Civ. R. 56.1(a) (emphasis added). That is not what Plaintiffs did. Plaintiffs submitted 940 separately numbered paragraphs spanning nearly 160 pages. Such a voluminous submission does not comply with the local rule. *Tripathy v. McCloskey*, No. 21-CV-6584 (CS), 2024 WL 2135623, at *1 (S.D.N.Y. May 13, 2024) (striking plaintiff's 356-page submission for "ignor[ing] the Rule's requirement that a statement be short and concise"). Plaintiffs' approach has made it extremely difficult for Defendants to respond fully while also providing their responses in a format useful to the Court.

To deal with this massive and improper Statement, Defendants have attempted to streamline their response for the benefit of the Court as follows:

Undisputed/Disputed/Disputed in Part. Every fact includes one of these responses, to make clear Defendants' position under Local Rule 56.1(b). Facts designated "Undisputed" are facts that Defendants have elected not to dispute solely for purposes of Plaintiffs' Motion for Summary Judgment in this matter, but does not constitute a binding admission for any other purpose. Facts designated "Disputed" or "Disputed in Part" are facts that Defendants are disputing.

In addition, Defendants have included certain objections and other responses to Plaintiffs' facts where appropriate. Those additional responses and the authority supporting those responses are identified below. For the Court's convenience and to avoid further lengthening this response, Defendants do not repeat their supporting authority for these issues in every response. Instead, Defendants state these objections plainly for ease of review and incorporate the below authority into each of those responses. Where warranted and helpful for the Court, Defendants also provide additional, paragraph-specific information to assist the Court's assessment.

Immaterial. Facts designated "Immaterial" are facts that, even if true, do not help resolve any question currently before the Court on summary judgment, i.e., they do not "affect the outcome of the suit under the governing law." *I.M. v. United States*, 362 F. Supp. 3d 161, 190 (S.D.N.Y. 2019). Rule 56.1 statements are "immaterial" if they "do not provide any basis to rule in favor of either party for summary judgment even if they are disputed." *See Parks v. Lebhar-Friedman, Inc.*, No. 04-cv-7133-DCP-KNF, 2008 WL 3833802, at *4 (S.D.N.Y. Aug. 11, 2008). Facts designated as "Immaterial" below thus reflect facts that are not legally significant because it does not impact an element of any claim or defense at issue in Plaintiffs' summary judgment motion. Defendants' designation of certain facts as "immaterial" does not concede that any fact not designated as such is material for purposes of adjudicating Plaintiffs' Motion, or for any other purpose.

Not Supported by Cited Evidence. Facts designated as "Not Supported by Cited Evidence" are facts with incorrect citations or where the cited evidence does not support the stated fact. *Holtz v. Rockefeller & Co.*, 258 F.3d 62, 74 (2d Cir. 2001) (A "Rule 56.1 statement is not itself a vehicle for making factual assertions that are otherwise unsupported in

the record. Where … the record does not support the assertions in a Local Rule 56.1 statement, those assertions should be disregarded and the record reviewed independently.").

Not Supported by Admissible Evidence. Facts designated as "Not Supported by Admissible Evidence" are facts that are not supported by admissible evidence as required by Local Rule 56.1, and therefore are not properly before the Court. S.D.N.Y. Local Rule 56.1(d) ("Each statement by the movant … must be followed by citation to evidence that would be admissible[.]"); *Epstein v. Kemper Ins. Companies*, 210 F. Supp. 2d 308, 314 (S.D.N.Y. 2002) (statements in a Rule 56.1 statement "are inappropriate if they are not based on personal knowledge, contain inadmissible hearsay, … or do not cite to supporting evidence").

Improper Legal Argument. Facts designated as an "Improper Legal Argument" are facts where the assertion therein constitutes improper attorney argument and/or is conclusory. These facts require no response. *Id.* (statements in a Rule 56.1 statement "are inappropriate if they are conclusory or argumentative[.]"); *In re LIBOR-Based Fin. Instruments Antitrust Litig.*, 801 F. Supp. 3d 330, 369 (S.D.N.Y. 2025) (quoting *Kesner v. Buhl*, 590 F. Supp. 3d 680, 691 (S.D.N.Y. 2022)) (the Court sets aside "[t]he portions of [a responding party's] 56.1 Statement that contain legal argument bereft of factual matter[.]").

Incomplete. Facts designated as "Incomplete" are facts where Plaintiffs have "introduce[d] all or part of a statement" and Defendants introduce "any other part—or any other statement—that in fairness ought to be considered at the same time" pursuant to Federal Rule of Evidence 106.

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.	1. **Undisputed.**
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.	2. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Appendix A identifies the works that Book Plaintiffs designate as the "Asserted Works" and purports to provide information concerning those works, including author, copyright holder, copyright registration number, first publication date, effective registration date, and publishing royalty recipient. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. Further, the Court should disregard the disputed portion of Paragraph 2 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.
3. The copyright claimant registered each Asserted Work with the United States Copyright Office within five years of its first publication. Appx. A.	3. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Appendix A purports to identify registration dates for

	each Asserted Work. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 3 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.
4. Each Asserted Work is currently available for sale. Appx. A.	4. **Undisputed.**

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.	5. **Undisputed.**
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.*	6. **Undisputed.**
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.	7. **Undisputed.**

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.	8. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Columbus Rose, Ltd. is identified in Appendix A as the copyright holder of 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. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 8 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

2. <u>Taylor Branch</u>

9. Plaintiff Taylor Branch is an author and historian who resides in Maryland. Nelson Decl. Ex. 69.	9. **Undisputed.**
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.*	10. **Undisputed.**
11. Branch has also received a MacArthur Fellowship and the National Humanities Medal. *Id.*	11. **Undisputed.**
12. Taylor Branch owns the valid copyrights in two published books: *Parting the Waters: America in the*	12. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed

King Years, 1954-63 and *Pillar of Fire: America in the King Years, 1963-65*. Appx. A.	that Taylor Branch is identified in Appendix A as the copyright holder of two published books: *Parting the Waters: America in the King Years, 1954-63* and *Pillar of Fire: America in the King Years, 1963-65*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 12 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

3. Michael Connelly and Hieronymus Inc.

13. Plaintiff Michael Connelly is an author who resides in Florida. Nelson Decl. Ex. 110.	13. **Undisputed.**
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.*	14. **Undisputed.**
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.*	15. **Undisputed.**
16. Plaintiff Hieronymus Inc. is a corporation through which Michael Connelly conducts his professional	16. **Undisputed.**

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.	17. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Hieronymus Inc. is identified in Appendix A as the copyright holder of thirty-six published books authored by Michael Connelly, including the *Harry Bosch* series, the *Mickey Haller* (Lincoln Lawyer) series, and the *Renee Ballard* series. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 17 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

4. Sylvia Day and Sylvia Day LLC

18. Plaintiff Sylvia Day is an author who resides in Nevada. Nelson Decl. Ex. 124.	18. **Undisputed.**
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.*	19. **Undisputed.**
20. Plaintiff Sylvia Day LLC is a limited liability company through which Sylvia Day conducts her professional activities, including the ownership and	20. **Undisputed.**

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.	21. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Sylvia Day LLC and Sylvia Day are identified in Appendix A as the copyright holders of twenty-nine published books authored by Sylvia Day, including the *Crossfire* series, which begins with *Bared to You*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 21 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

5. Jonathan Franzen

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

25. Jonathan Franzen owns the valid copyrights in five published books, including *The Corrections*, *Freedom*, and *Crossroads*. Appx. A.	25. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Jonathan Franzen is identified in Appendix A as the copyright holder of five published books, including *The Corrections*, *Freedom*, and *Crossroads*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 25 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

6. Christopher Golden and Daring Greatly Corporation

26. Plaintiff Christopher Golden is an author who resides in Massachusetts. Nelson Decl. Ex. 103.	26. **Undisputed.**
27. Golden is a *New York Times* bestselling horror and thriller novelist who has been called the "the king of horror-thriller." *Id.*	27. **Undisputed.**
28. His work has been published in dozens of languages, and he is the winner of the Bram Stoker Award. *Id.*	28. **Undisputed.**
29. Plaintiff Daring Greatly Corporation is a Massachusetts-based corporation through which Christopher Golden conducts his professional activities,	29. **Undisputed.**

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.	30. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Christopher Golden is identified in Appendix A as the copyright holder of one published book: *Ararat*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 30 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

7. Andrew Sean Greer

31. Plaintiff Andrew Sean Greer is an author who resides in California. Nelson Decl. Ex. 70.	31. **Undisputed.**
32. Greer's *Less* won the 2018 Pulitzer Prize for Fiction and was a national bestseller. *Id.*	32. **Undisputed.**
33. Greer has also received a Guggenheim Fellowship, the California Book Award, and the New York Public Library Young Lions Award. *Id.*	33. **Undisputed.**
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.	34. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Andrew Sean Greer is identified in Appendix A as the copyright holder of four published books, including *Less*,

	Less Is Lost, *The Confessions of Max Tivoli*, and *The Story of a Marriage*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 34 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

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

35. Plaintiff John Grisham is an author who resides in Virginia. Nelson Decl. Ex. 85.	35. **Undisputed.**
36. Grisham has published more than fifty consecutive number one bestsellers, and his books have been translated into nearly fifty languages. *Id.*	36. **Undisputed.**
37. He has twice won the Harper Lee Prize for Legal Fiction and received the Library of Congress Creative Achievement Award for Fiction. *Id.*	37. **Undisputed.**
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.	38. **Undisputed.**
39. Belfry Holdings, Inc. and John Grisham own the valid copyrights in twenty-three published books	39. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed

authored by John Grisham, including *A Time to Kill*, *The Firm*, *The Pelican Brief*, and *The Client*. Appx. A.	that Belfry Holdings, Inc. and John Grisham are identified in Appendix A as the copyright holders of twenty-three published books authored by John Grisham, including *A Time to Kill*, *The Firm*, *The Pelican Brief*, and *The Client*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 39 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

9. David Henry Hwang

40. Plaintiff David Henry Hwang is an author and playwright who resides in New York. Nelson Decl. Ex. 793.	40. **Undisputed.**
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.	41. **Undisputed.**
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.	42. **Undisputed.**
43. David Henry Hwang owns the valid copyrights in three published books, including *M. Butterfly*. Appx. A.	43. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that David Henry Hwang is identified in Appendix A as the

	copyright holder of three published books, including *M. Butterfly*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 43 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

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

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

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.	48. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Wo & Shade, LLC is identified in Appendix A as the copyright holder of fifteen published books authored by George R.R. Martin, including novels in the *A Song of Ice and Fire* series. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 48 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

11. <u>Jodi Picoult</u>

49. Plaintiff Jodi Picoult is an author who resides in New Hampshire. Nelson Decl. Ex. 86.	49. **Undisputed.**
50. Picoult has written twenty-nine novels translated into forty languages, with an estimated 63.7 million copies in print. *Id.*	50. **Undisputed.**
51. Her honors include the New England Bookseller Award for Fiction and the Sarah Josepha Hale Award. *Id.*	51. **Undisputed.**
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.	52. **Disputed in Part – Improper Legal Argument; Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that Jodi Picoult is identified in Appendix A as the copyright holder

	of published books including *My Sister's Keeper*, *Nineteen Minutes*, and *Small Great Things*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 52 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Further, the disputed proposition is not supported by the cited evidence. Appendix A lists only twenty-five, not twenty-six, works attributed to Picoult.

12. Stacy Schiff and Gattopardo, LLC

53. Plaintiff Stacy Schiff is an author who resides in New York. Nelson Decl. Ex. 95.	53. **Undisputed.**
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.*	54. **Undisputed.**
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.*	55. **Undisputed.**
56. Plaintiff Gattopardo, LLC is a limited liability company through which Stacy Schiff conducts her	56. **Undisputed.**

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.	57. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Stacy Schiff is identified in Appendix A as the copyright holder of three published books, including *The Witches and Cleopatra: A Life*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 57 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

13. <u>James Shapiro</u>

58. Plaintiff James Shapiro is an author who resides in New York. Nelson Decl. Ex. 82.	58. **Undisputed.**
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.	59. **Undisputed.**
60. *A Year in the Life of William Shakespeare: 1599* won the Samuel Johnson Prize, and *The Year of Lear*	60. **Undisputed.**

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.	61. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that James Shapiro is identified in Appendix A as the copyright holder of three published books, including *A Year in the Life of William Shakespeare: 1599* and *The Year of Lear: Shakespeare in 1606*. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. The Court should disregard the disputed portion of Paragraph 61 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response.

14. The Authors Guild (Estate of Mignon Eberhart)

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.	62. **Undisputed.**
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.	63. **Undisputed.**

64. Mignon Eberhart was an author who published more than fifty mystery novels over a sixty-year career. Nelson Decl. Ex. 87.	64. **Undisputed.**
65. She is a former president of the Mystery Writers of America and a recipient of the organization's Grand Master Award. *Id.*	65. **Undisputed.**
66. Eberhart bequeathed the copyrights in her published books to the Authors Guild, which acquired those copyrights upon her death. Nelson Decl. Ex. 136.	66. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that Eberhart bequeathed the copyrights in her published books to the Authors Guild, which acquired those copyrights upon her death. However, the Court should disregard the disputed portion of Paragraph 66 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions.
67. The Authors Guild owns the valid copyrights in five published books authored by Mignon Eberhart. Appx. A.	67. **Disputed in Part – Improper Legal Argument; Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that the Authors Guild claims to own copyrights in published books authored by Mignon Eberhart. However, disputed to the extent Plaintiffs seek to establish any presumption of validity or ownership of any Asserted Work under 17 U.S.C. § 410(c), because those are improper legal conclusions. Further, the Court should disregard the disputed portion of Paragraph 67 because the assertion constitutes improper

	attorney argument and/or is conclusory. It thus requires no response. Further, the disputed proposition is not supported by the cited evidence. Plaintiffs' cited evidence also does not establish ownership and validity as to all five works. Although Plaintiffs [redacted], the cited evidence does not establish [redacted]. For *Hasty Wedding* and *Escape the Night*, Plaintiffs have not cited copyright renewal registrations. Additionally, Plaintiffs' witness, Ms. Mary Rasenberger, testified that [redacted]. C. Blythe Opp. Decl. Ex. 1 (M. Rasenberger 30(b)(6) Tr. (Feb. 13, 2026)) 100:23–101:7. *See also* Books Ex. 136 at FICTION-0000004-009.

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.	68. **Undisputed.**
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.	69. **Undisputed.**

70. Defendant OpenAI GP, LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 45, ECF No. 754.	70. **Undisputed.**
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.	71. **Undisputed.**
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.	72. **Undisputed.**
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.	73. **Undisputed.**
74. Defendant OpenAI, LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 47, ECF No. 754.	74. **Undisputed.**
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.	75. **Undisputed.**
76. Defendant OpenAI Global LLC is a Delaware limited liability company headquartered in California. OpenAI's Answer ¶ 48, ECF No. 754.	76. **Undisputed.**
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.	77. **Undisputed.**

78. Defendant OpenAI Holdings, LLC is a Delaware limited liability company. OpenAI's Answer ¶ 50, ECF No. 754.	78. **Undisputed.**
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.	79. **Undisputed.**
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.	80. **Undisputed.**
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.	81. **Undisputed.** Defendants respond as follows: Undisputed as to the asserted facts; Defendants note, however, that the Consolidated Amended Complaint appears at ECF 183 and not ECF 138 as Plaintiffs represent.
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.	82. **Undisputed.** Defendants respond as follows: Undisputed as to the asserted facts; Defendants note, however, that the Consolidated Amended Complaint appears at ECF 183 and not ECF 138 as Plaintiffs represent.
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.	83. **Undisputed.** Defendants respond as follows: Undisputed as to the asserted facts; Defendants note, however, that the Consolidated Amended Complaint appears at ECF 183 and not ECF 138 as Plaintiffs represent.

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."	84. **Disputed.** Defendants respond as follows: Defendants dispute the use of the term "Defendants" to the extent it is used in any statement herein to refer to only OpenAI or only Microsoft, and not both.

C. OpenAI's Conversion from Nonprofit to Commercial Enterprise

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.	85. **Undisputed.** Microsoft defers to OpenAI.
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.	86. **Disputed in Part.** OpenAI responds as follows: Undisputed that by 2017, OpenAI's founders were considering an organizational change that would allow investment rather than donations alone, and that 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." However, disputed to the extent Plaintiffs suggest that OpenAI ceased to have a nonprofit component. The cited testimony reflects that OpenAI retained its nonprofit organization following restructuring.

	OpenAI further disputes Plaintiffs' characterization of the reason for the restructuring as incomplete and misleading. The testimony explains that OpenAI's increasing computing needs led it to adopt a structure capable of attracting the investment necessary to support those needs. *See* Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 171:17–24. Microsoft defers to OpenAI.
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.*	87. **Disputed in Part.** OpenAI responds as follows: Undisputed that in March 2019, OpenAI formed OpenAI LP, which could issue equity to raise capital and took over the majority of OpenAI's operations. However, disputed that Plaintiffs' characterization of OpenAI LP as merely a "for-profit subsidiary" is accurate or complete. OpenAI LP was structured as a "capped-profit" entity, a hybrid model designed to balance mission-driven research with the need to attract investment. *See* Books Ex. 706. Microsoft defers to OpenAI.
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.	88. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 365 contains the quoted text. Disputed that Plaintiffs' misleading characterization of, and inferences drawn from, Books Ex. 365 establish the reasons for OpenAI's organizational restructuring. Other portions of the same contemporaneous journal in Books Ex. 365 reflect a broader purpose, including that "our point isn't making the discoveries, it's ensuring that it's a public good." Consistent with that context, Brockman testified that

	OpenAI came to believe that "a for-profit entity of some form would be really the only way to successfully pursue the mission." C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 450:3–6. Further disputed that this passage from Brockman's personal journal is relevant to the claims or defenses at issue, and OpenAI objects under Fed. R. Evid. 401 and 403. Disputed as immaterial for the same reason and because this fact does not affect the outcome of the suit under the governing law. Microsoft defers to OpenAI.
89. Brockman also wrote that he "was deeply motivated by the gazillions." Nelson Decl. Ex. 367 at -624.	89. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 367 contains the quoted text. Disputed to the extent the statement is attributed to Brockman or to OpenAI more generally. Books Ex. 367 contains statements from multiple individuals. Further disputed that Brockman's personal journal and Plaintiffs' interpretation of Brockman's personal motivations are relevant to the claims or defenses at issue, and OpenAI objects under Fed. R. Evid. 401 and 403. Further disputed as immaterial for the same reason and because this fact does not affect the outcome of the suit under the governing law. Microsoft defers to OpenAI.

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.	90. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 366 contains the quoted text. However, Plaintiffs' characterization of, and inferences drawn from Ex. 366 are disputed. Plaintiffs take the quote out of its context, which included a list of several company goals that included: "preserve liberal values in US & other countries that already have them," "distribute the vast benefits of AGI equally," and "spread liberal values around the world." Books Ex. 366 at OPCO_MDL_PD_TEMPBATES_00618-19. Further disputed that Brockman's personal journal and Plaintiffs' interpretation of Brockman's personal motivations are relevant to the claims or defenses at issue, and OpenAI objects under Fed. R. Evid. 401 and 403. Further disputed as immaterial for the same reason and because this fact does not affect the outcome of the suit under the governing law. Microsoft defers to OpenAI.
91. OpenAI's valuation has since reached at least $852 billion. Nelson Decl. Ex. 707.	91. **Undisputed.** Microsoft defers to OpenAI.
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.	92. **Undisputed.** Microsoft defers to OpenAI.

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.	93. **Disputed.** OpenAI responds as follows: Disputed to the extent Plaintiffs incorporate by reference the facts and characterizations set forth in Section I.E of their statement of undisputed facts. OpenAI incorporates by reference its responses and objections to the facts set forth in Section I.E. Microsoft defers to OpenAI.

D. Generative AI and Large Language Models

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.	94. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** Defendants respond as follows: Defendants do not dispute that large language models ("LLMs") are a kind of "generative artificial intelligence," that generates outputs, such as text and images. But Defendants dispute the characterization that LLM outputs are based solely on the training data. Rather, LLMs generate outputs based on user inputs, context, a network of up to trillions of numerical parameters (mathematical weights) that encode generalized statistical relationships and patterns (such as syntax, semantics, and factual associations), derived from the entire training corpus, that it can use to generate new text—i.e., to generalize broader ideas and principles of language. *See* ECF 1723 (J. Lafferty Decl.) ¶¶ 16, 33-35, 45, 52, 56, 64-66, 71-72, 74-77, 80-81, 90-91, 94-100, 115-118, 127-134, 139-41, 154-55, 157-58, 164, 174, 176.

	Defendants also dispute Plaintiffs' characterization of training data as the "inputs" for which a generative AI system generates outputs accurately reflects the cited authority. Plaintiffs' cited authority instead explains that generative AI models create new data samples in response to user input data based on learned patterns from the data upon which the model was trained. *See* Books Ex. 679 at 112 ("A generative AI model is a type of machine learning architecture that uses AI algorithms to create novel data instances, drawing upon the patterns and relationships observed in the training data."); *id*. ("Building upon the above, a generative AI model refers to generative modeling that is instantiated with a machine learning architecture (e.g., a deep neural network) and, therefore, can create new data samples based on learned patterns."). Defendants also respond that the facts set forth in Paragraph 94 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Books Ex. 679 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability under Fed. R. Evid. 803(18), and therefore may not be considered in support of summary judgment.

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.	95. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that pre-training is part of the LLM training process. Undisputed that the models at issue underwent pre-training. However, disputed that Plaintiffs' characterization of pre-training as the "core" of the LLM training process is supported by the cited evidence. The cited evidence does not characterize pre-training as the "core" of that process. *See* Books Ex. 680. Further, Plaintiffs' characterization is inaccurate because post-training is what provides LLMs their defining features and capabilities. *See generally* ECF 1748 (C. Callison-Burch Decl.) Section VIII.
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*.	96. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that Books Ex. 681 provides a description of pretraining. However, disputed that the cited document supports Plaintiffs' characterization of the pre-training process as presenting a model with "incomplete text data, such as part of a sentence" and then having the model predict the next word or fragment of a word. The disputed proposition is not supported by the cited evidence. Specifically, the cited documents do not support Plaintiffs' assertion that "[a]t the start of this pre-training process, outputs are effectively random." Neither Exhibit 681 nor Exhibit 680 make such characterizations. *See* Books Ex. 681; Books Ex. 680.

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.	97. **Disputed – Not Supported by Cited Evidence & Cited Evidence is Inadmissible.** Defendants respond as follows: The facts set forth in Paragraph 97 are disputed because they are not supported by the cited evidence. Specifically, disputed that Books Ex. 700 supports Plaintiffs' characterization of the training models at issue. The cited paper, published in 2015, does not address unsupervised pre-training of transformer-based LLMs such as the GPT models at issue. Further, Plaintiffs' characterization that a model's prediction is "checked against the training data" does not accurately reflect the cited paper, which more precisely describes comparing model output against a target output. *See id.* Defendants also object that the cited evidence is inadmissible hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability under Fed. R. Evid. 803(18), and therefore may not be considered in support of summary judgment.
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.	98. **Disputed – Not Supported by Cited Evidence & Cited Evidence is Inadmissible.** Defendants respond as follows: The facts set forth in Paragraph 98 are disputed because they are not supported by the cited evidence. Specifically, disputed that Books Ex. 700 supports Plaintiffs' characterization of the training models at issue. The cited paper, published in 2015, does

	not address unsupervised pre-training of transformer-based LLMs such as the GPT models at issue. Further, the cited paper does not support Plaintiffs' characterization of "loss" as the measure of how much a model's prediction "varies from the actual data in its training data"; the cited paper instead discusses an "error" measured through an objective function. Additionally, the cited paper does not use the terms "weights," "parameters," and "loss" consistently with how OpenAI uses these terms when training its models. *See id.* Defendants also object that the cited evidence is inadmissible hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability under Fed. R. Evid. 803(18), and therefore may not be considered in support of summary judgment.
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.").	99. **Undisputed.** Microsoft defers to OpenAI.
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.	100. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 100 are disputed because they are not supported by the cited evidence. Specifically, the cited testimony explains that the goal of pretraining is to predict

	a probability distribution, *not* to "match the training data as closely as possible." OpenAI further responds that it disputes the premise in Paragraph 100 that using next-token prediction loss to pretrain an LLM means "train[ing] the model so that its outputs match the training data as closely as possible." Rather, pretraining is meant to build a model capable of generalization, not to "minimize the loss." Truly minimizing train loss would lead to "overfitting," something OpenAI works hard to avoid. *See* C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr.) at 175:6–21, 178:9–14; *see also* ECF 1748 (C. Callison-Burch Decl.) ¶¶ 45–50 (discussing "test loss" vs. "train loss"). Microsoft defers to OpenAI.
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).	101. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 298 contains the text quoted in Plaintiffs' cited support. However, disputed that the cited evidence supports Plaintiffs' characterization that the "core function" of an LLM is to generate text similar to its training data. Rather, the goal of an LLM is generalization. See, e.g., ECF 1748 (C. Callison-Burch Decl.) ¶¶ 65–66; Books Ex. 751 ¶¶ 268-271. Further disputed that the cited slide characterizes regurgitation as an LLM's "core function"; the slide instead identifies regurgitation as a problem to be

	addressed or mitigated. *See* Books Ex. 298. OpenAI also incorporates by reference its response to Fact 100. Microsoft defers to OpenAI.
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.	102. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 709 contains the quoted text. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Books Ex. 709 establish that Betker was describing large language models (LLMs) when he stated that generative AI models "truly approximate their datasets to an incredible degree." Specifically, disputed that the cited document concerns LLMs; Betker testified that the document was primarily about the image-generation model DALL-E 3 and that he had never trained a language model. *See* C. Blythe Opp. Decl. Ex. 3 (J. Betker 30(b)(1) Tr.) at 93:13-95:4, 115:8-14; Books Ex. 709. OpenAI also incorporates by reference its response to Fact 100. Microsoft defers to OpenAI.
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	103. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 709 contains the quoted text. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Books Ex. 709 establish that Betker was describing LLMs or text-generation models. Rather, as noted in response to Fact 102, Betker testified that the document was primarily about the image-generation model

"'Lambda,' 'ChatGPT,' 'Bard,' or 'Claude,' . . . it's not the model weights that you are referring to. It's the dataset." *Id.*	DALL-E 3 and that he had never trained a language model. *See* C. Blythe Opp. Decl. Ex. 3 (J. Betker 30(b)(1) Tr.) at 93:13-95:4, 115:8-14; Books Ex. 709. Further disputed that the bracketed insertion "[technical choices around model training such as]" accurately characterizes or is supported by Betker's statement. *See* Books Ex. 709. The use of "such as" in the inserted language broadens the scope of the statement beyond what Betker initially wrote and overgeneralizes Betker's specific enumeration of three discrete technical variables into a single broad, undefined category. OpenAI also incorporates by reference its response to Fact 100. Microsoft defers to OpenAI.

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.	104. **Undisputed.**
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.	105. **Undisputed.** Microsoft defers to OpenAI.
106. That interface was OpenAI's first commercial product, and OpenAI described it as a "revenue source to	106. **Undisputed.**

help [the company] cover costs in pursuit of [its] mission." Nelson Decl. Ex. 720.	Microsoft defers to OpenAI.
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.	107. **Undisputed.** Microsoft defers to OpenAI.
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.	108. **Undisputed.** Microsoft defers to OpenAI.
109. ChatGPT reached 100 million monthly active users within two months—the fastest-growing consumer internet application in history. Nelson Decl. Ex. 721.	109. **Undisputed.** Microsoft defers to OpenAI.
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.	110. **Disputed in Part.** OpenAI responds as follows: Undisputed that OpenAI launched ChatGPT Enterprise in August 2023 and ChatGPT Team in January 2024. However, disputed that the cited evidence independently establishes the stated launch dates and product details, as Plaintiffs rely solely on M. Lasinski Decl. Ex. A at ¶ 43, an expert declaration, for these assertions. Disputed that ChatGPT Plus is properly characterized merely as a "$20-per-month subscription" to the extent that characterization implies that OpenAI's purpose in launching ChatGPT Plus was principally to monetize ChatGPT. Turley testified that the ChatGPT research demo had experienced substantial demand and downtime, and that OpenAI explored subscriptions "as a solution to provide

	uptime," explaining that he "would not have described the goal of that effort to make money," but rather to provide "a way of accessing the demo without downtime." C. Blythe Opp. Decl. Ex. 4 (N. Turley 30(b)(6) Tr.) at 92:11–22. Microsoft defers to OpenAI.
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.	111. **Disputed in Part.** OpenAI responds as follows: Undisputed that 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. However, disputed that the cited evidence independently establishes these release dates, as Plaintiffs rely solely on ECF 1825-1 (M. Lasinski Decl. Ex. A) ¶¶ 41, 43, an expert declaration, for these assertions. Further disputed that each identified model became generally available through both ChatGPT and OpenAI's application programming interface at the time of its stated release, as the cited evidence does not establish when each model became available through each interface. For example, access to the GPT-4 API was initially limited by a waitlist rather than made generally available to users upon release. C. Blythe Opp. Decl. Ex. 5 at OPCO_MDL_001934008. Microsoft defers to OpenAI.
112. OpenAI trained GPT-4 in late 2021 and 2022 and released it on March 14, 2023. Lasinski Decl. Ex. A at ¶ 41.	112. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that GPT-4 was released on March 14, 2023. However, The Plaintiffs' assertion that OpenAI trained

	GPT-4 "in late 2021 and 2022" is disputed because it is not supported by the cited evidence. Specifically, OpenAI's corporate representative testified that GPT-4 training began in February 2022, not in late 2021, and that GPT-4 training was completed in August 2022. C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) 36:18–37:2. Microsoft defers to OpenAI.
113. OpenAI retired GPT-4 in April 2025 and removed GPT-4o from ChatGPT in February 2026. Lasinski Decl. Ex. A at ¶ 42.	113. **Disputed in Part.** OpenAI responds as follows: Undisputed that GPT-4 was retired from ChatGPT in April 2025 and that GPT-4o was retired from ChatGPT in February 2026. However, disputed to the extent Plaintiffs' characterization suggests that OpenAI retired those models entirely or that they are no longer available to users. Both GPT-4 and GPT-4o have remained available through OpenAI's API following their retirement from ChatGPT. C. Blythe Opp. Decl. Ex. 7 (ChatGPT Release Notes) at 50, 86; ECF 1825-1 (M. Lasinski Decl. Ex. A) ¶ 42 nn.141, 142. Microsoft defers to OpenAI.
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.	114. **Undisputed.** Microsoft defers to OpenAI.
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	115. **Undisputed.** Microsoft defers to OpenAI.

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.	116. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that OpenAI generated revenue between 2019 and 2024 relating to ChatGPT and API products through which it offered certain of the Accused Models. However, disputed that Plaintiffs' assertion that OpenAI generated approximately [redacted] billion in revenue relating to the models at issue and the products through which OpenAI offered them accurately reflects the cited evidence, as the revenue figures reflected in Books Ex. 353 total approximately [redacted] billion when the identified ChatGPT and API revenue amounts are combined, not [redacted] billion. Disputed to the extent Plaintiffs characterize that revenue as relating to "the products through which [OpenAI] offered [the models at issue]," because that phrase is undefined and the cited evidence does not attribute all of the identified revenue to particular models at issue. Further, the disputed proposition is not supported by the cited evidence. Books Ex. 353. Microsoft defers to OpenAI.
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]	117. **Disputed in Part – Immaterial.** OpenAI responds as follows: Undisputed that Books Ex. 326 contains the stated historical revenue and gross-profit figures for 2023 and 2024 and the stated estimates for 2025.

[redacted] in 2023, [redacted] in 2024, and an estimated [redacted] in 2025. *Id.*	However, disputed to the extent Plaintiffs characterize OpenAI's 2025 forecast of $12.7 billion in revenue and estimated [redacted] in gross profit as actual financial results; those figures are projections, not realized revenue or gross profit. Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, these company-wide financial figures suggest that they reflect revenue or gross profit attributable to the models at issue; the cited evidence does not establish such an attribution. The disputed proposition in Paragraph 117 is therefore immaterial because it does not affect the outcome of the suit under the governing law. Microsoft defers to OpenAI.
118. ChatGPT and the OpenAI application programming interface [redacted] of OpenAI's revenues and gross profits in 2023 and 2024, and for [redacted] in 2025. Nelson Decl. Ex. 326 at -677.	118. **Disputed in Part – Immaterial.** OpenAI responds as follows: Undisputed that Books Ex. 326 contains the stated figures for 2023 and 2024 and the stated estimate for 2025. However, disputed to the extent Plaintiffs characterize the figures for 2025 as reflecting actual revenue, as it is based on projected, not realized, 2025 revenue. Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, company-wide revenue and gross-profit figures suggest that they are attributable to the models at issue; the cited evidence does not establish such an attribution. The disputed proposition in Paragraph 118 is therefore immaterial because it does not affect the outcome of the suit under the governing law.

	Microsoft defers to OpenAI.
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.	119. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the revenue figures asserted in Paragraph 119 are reflected in Books Ex. 353. However, disputed that the cited evidence supports Plaintiffs' assertion that [redacted] of the API revenue is attributable to the models at issue, as Books Ex. 353 does not identify or explain that figure or provide a basis for attributing that amount to the models at issue. Further disputed to the extent Plaintiffs' reference to "paid versions of ChatGPT" suggests that the cited revenue can be attributed to particular models at issue, as the cited evidence does not establish that attribution. For those reasons, the disputed propositions are not supported by the cited evidence. Books Ex. 353. Microsoft defers to OpenAI.
120. OpenAI also received approximately [redacted] parties' [redacted] Nelson Decl. Ex. 213.	120. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Defendants do not dispute that OpenAI received approximately [redacted] from Microsoft under the parties' [redacted] from January 2023 through December 2024. However, Defendants dispute that the cited evidence supports Plaintiffs' assertion that OpenAI received approximately [redacted] from Microsoft. Specifically, Books Ex. 213 reflects [redacted] Defendants further

	dispute this fact to the extent Plaintiffs' characterization suggests that the asserted [redacted] is attributable to the models at issue; the cited evidence does not establish such an attribution. Books Ex. 213.

F. Microsoft

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.	121. **Disputed in Part – Not Supported by Cited Evidence.** Microsoft does not dispute that it is a Washington-based corporation that develops and sells software, cloud computing services, and products employing artificial intelligence technology. Defendants do not dispute that OpenAI used Azure infrastructure to train and deploy GPT-3 and later models. Defendants do not dispute that storage space is needed to store any form of data, including training data, and that Microsoft provided OpenAI with storage space (called Blob) to utilize as needed. But Plaintiffs' cited evidence does not establish that the identified storage was "critical" to training the at-issue LLMs as Plaintiffs rely solely on deposition testimony from a Microsoft employee (Mr. Melton) who lacks personal knowledge of OpenAI's training infrastructure and operations. Pursuant to Fed. R. Evid. 106, Mr. Melton stated that "Blob Storage is not exclusive to OpenAI, Blob is used for everything, right, that's the method of storage." *See* Books Ex. 36 (L. Melton 30(b)(6) Tr.) at 61:4-63:7.

	Defendants further dispute this fact as unsupported as the cited evidence says nothing about "Microsoft's training data", what that data is, or any fact regarding what "OpenAI understood." The fact misquotes the cited evidence and makes false statements regarding "Microsoft's training data" and "OpenAI's understanding" regarding that data. *See* Books Ex. 36 at 61:4-63:7.
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.	122. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** Defendants do not dispute that Microsoft and OpenAI entered a strategic partnership in July 2019, which is memorialized in the first Joint Development and Collaboration Agreement. But Defendants respond that the agreement cited as Books Ex. 145 is not the agreement pursuant to which Microsoft committed $1 billion in capital. Rather, a separate agreement—the [redacted]—contains Microsoft's commitment of $1 billion in capital (as a combination of Azure compute credits and cash), [redacted]. That agreement is not cited. Further disputed to the extent Plaintiffs characterize Microsoft as OpenAI's "exclusive cloud provider." The JDCA required OpenAI, Inc. and OpenAI, L.P. to [redacted] Books Ex. 145. Finally, Books Ex. 203 is a draft Microsoft memo that does not identify the author or recipient ("To:" and "From:" are blank), and

	therefore cannot be substantiated as containing correct information and is thus not admissible evidence, as required by Local Rule 56.1.
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.	123. **Undisputed.**
124. In September 2020, Microsoft obtained an exclusive license to GPT-3. Lasinski Decl. Ex. A at ¶ 73.	124. **Disputed in Part.** Defendants respond as follows: Undisputed that Microsoft obtained a license to GPT-3 in September 2020 for their own products and services under which Microsoft was the only third party granted the licensed usage rights. However, disputed to the extent Plaintiffs' characterization of the license as "exclusive" suggests that OpenAI relinquished its own rights to use GPT-3. OpenAI retained its own use rights, and the agreement did not affect continued access to GPT-3 through OpenAI's API. See ECF 1825-1 (M. Lasinski Decl. Ex. A) ¶ 73, n. 351.
125. Microsoft receives 20 percent of OpenAI's revenue, held an approximately 27 percent 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.	125. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Defendants do not dispute that Microsoft receives a revenue share for a portion of OpenAI's revenue, that Microsoft "holds approximately 27 percent in the PBC on an as-converted diluted basis…and Microsoft will no longer have a right of first refusal to be OpenAI's compute provider" as of September 2025. Books Ex. 736 at 27. Microsoft does not dispute that it has deployed OpenAI's GPT models across at least [redacted]

	Microsoft products. Defendants dispute that Microsoft receives 20 percent of OpenAI's total revenue, as the cited agreement provides that Microsoft would receive 20% of "Net Revenue." Books Ex. 146(a) at MSFT_AI_MDL_000000581 (defining "Net Revenue"); Books Ex. 146(b) at MSFT_AI_MDL_000000596. Further, the revenue that Microsoft receives from OpenAI is for only the OpenAI API and ChatGPT [redacted] Additionally, [redacted] C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 210:23-212:12; 214:1-215:24; Books Ex. 213. Microsoft also responds that any deployment of OpenAI's GPT models in Bing or the list of [redacted] products in the Thornburgh declarations is immaterial because this fact does not affect the outcome of the suit under the governing law. Neither Bing nor the list of [redacted] products in the Thornburgh declarations are accused products in this case. *See, e.g.* ECF 183 (Compl.), ¶¶ 5, 17-18. Additionally, Defendants dispute the implication that Microsoft has deployed all of OpenAI's GPT models across more than [redacted] Microsoft products because it is not supported by the cited evidence. The cited support only discusses implementations of GPT-4o models. ECF 858

	(Thornburgh Decl.) ¶ 8. The Thornburgh declaration is also incorrectly cited as ECF 585, and should be ECF 858.

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.	126. **Disputed – Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: Disputed that Plaintiffs' characterization of LibGen as a website that "collects, hosts, and distributes digital copies of copyrighted books without charge and without authorization" accurately reflects the cited evidence. Plaintiffs' cited materials describe LibGen more broadly as "a search engine for articles and books on various topics" and do not establish that all materials available through LibGen are copyrighted books or are available without authorization from the applicable rightsholders. *See* Books Ex. 384; Books Ex. 385. For those reasons, the disputed portion of Paragraph 126 is unsupported by the record. OpenAI also objects to Plaintiffs' reliance on Seeley's cited opinions, which Defendants have moved to exclude under Fed. R. Evid. 702. Microsoft defers to OpenAI.
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.*	127. **Disputed in Part.** OpenAI responds as follows: Undisputed as to the time and place of LibGen's origin. However, disputed that the cited evidence supports

	the assertion that LibGen incorporated books from earlier pirate libraries at the time of its inception. Specifically, the paragraph of Seeley's report that Plaintiffs rely upon observes only that LibGen "[redacted]" *See* ECF 1804-1 (Seeley Decl. Ex. A) ¶ 21. Seeley's speculation does not support Plaintiffs' conclusion that "[a]t its inception, [LibGen] incorporated copyrighted books from earlier pirate libraries." Microsoft defers to OpenAI.
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.	128. **Disputed – Immaterial; Not Supported by Cited Evidence; Improper Legal Argument.** Defendants respond that the facts set forth in Paragraph 128 are disputed. As it relates to Microsoft, the cited evidence does not support the statement. The cited evidence is deposition excerpts from OpenAI witnesses that do not mention Microsoft. Additionally, whether Microsoft's witnesses "understood that LibGen distributes large collections of unauthorized copies of copyrighted works" is immaterial to any claim asserted against Microsoft at issue in Book Plaintiffs' summary judgment motion. As it relates to OpenAI, disputed that Plaintiffs' characterization of what "Defendants' own witnesses understood" LibGen to distribute is supported by the cited

	evidence. Plaintiffs rely on testimony given in 2026 by two former OpenAI employees concerning [redacted]; that testimony does not establish what OpenAI, other OpenAI employees, or those witnesses [redacted]. C. Blythe Opp. Decl. Ex. 12 (B. Chess 30 (b)(1) Tr.) at 19:20–20:20, 21:1–4; C. Blythe Opp. Decl. Ex. 31 (L. Weng 30(b)(1) Tr.) at 22:14–23:12. Further disputed to the extent Plaintiffs state the two cited witnesses "understood that LibGen distributes large collections of unauthorized copies of copyrighted works," as this conclusion is not supported by the cited testimony and embeds a legal conclusion regarding whether copies are "authorized." Further disputed to the extent Plaintiffs' characterization suggests that LibGen distributes only "unauthorized copyrighted" books or similar works. Plaintiffs' own cited materials describe LibGen more broadly as "a search engine for articles and books on various topics" and do not establish that all materials available through LibGen are "copyrighted" or "unauthorized." Books Ex. 384 at WIKIPEDIA-OAI-00000002; Books Ex. 385 at WIKIPEDIA-OAI-00000003.
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.	129. **Disputed**. OpenAI responds as follows: Disputed that the cited evidence establishes that Alexandra Elbakyan submitted a declaration "on behalf of LibGen." Elbakyan was the operator of Sci-Hub and submitted a letter in the 2015 Elsevier litigation concerning Sci-Hub, and the cited evidence does not establish that she submitted a declaration on behalf of LibGen. Accordingly, Defendants

	dispute the claim that any purported "declaration" has ever been submitted "on behalf of LibGen." Further disputed that Elbakyan characterized LibGen's activities as "pirating." The statement that she "refer[red] to the websites' activities as 'pirating'" is the court's characterization of her submission, while Elbakyan's discussion of "pirating" concerned her own acquisition of academic research papers from unidentified websites when she lacked access to them while studying at Kazakhstan University. C. Blythe Opp. Decl. Ex. 9 (Letter from Pro Se Defendant Alexandra Elbakyan, *Elsevier, Inc. v. www.Sci-Hub.org*, No. 15-cv-04282 (S.D.N.Y. Sept. 15, 2015)). Microsoft defers to OpenAI.
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.	130. **Undisputed.** Microsoft defers to OpenAI.
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.	131. **Undisputed.** Microsoft defers to OpenAI.
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	132. **Undisputed.** Microsoft defers to OpenAI.

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.	133. **Undisputed.** Microsoft defers to OpenAI.
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.	134. **Undisputed.** Microsoft defers to OpenAI.
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.	135. **Undisputed.** Microsoft defers to OpenAI.
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	136. **Disputed in Part.** OpenAI responds as follows: Undisputed that LibGen generally divides its collection into fiction and nonfiction categories, assigns Md5 hashes, and lists metadata including title, author, ISBN, upload date, file format, and Md5 hash. However,

the files it collects, including title, author, ISBN, upload date, file format, and Md5 hash. *Id.*	disputed to the extent Plaintiffs claim LibGen's collection is accurately catalogued this way for all files. Further disputed to the extent Plaintiffs rely on LibGen metadata to establish the specific files OpenAI actually acquired. *See* S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 51-53, 65-69 (explaining that metadata-only identification does not compare underlying text and requires assumptions that OpenAI obtained the complete dataset, that the public metadata matches the materials OpenAI downloaded, and that the metadata is an accurate and complete catalog). Further, the metadata relied upon by Shan was downloaded in April 2024, years after the relevant 2019–2020 downloads, and therefore does not, without appropriate time filtering, establish the contents of LibGen at the relevant time. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 146-147. Microsoft defers to OpenAI.
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.	137. **Disputed in Part.** OpenAI responds as follows: Undisputed that as of October 2019, LibGen fiction metadata reflected 2.18 million entries and that, as of January 2020, LibGen nonfiction metadata reflected 2.46 million entries. However, disputed to the extent Plaintiffs imply that these metadata counts represent unique books or are evidence of material actually acquired by OpenAI. As explained by Frederiksen-Cross, LibGen metadata counts cannot simply be converted into a completed book-download count: many LibGen fiction records were duplicates, a queued torrent likely may not result in the

	download of all contained files, and the file sets included non-book content (poems, catalogs, pamphlets, academic papers, comics), and non-English works. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 141–149. Microsoft defers to OpenAI.

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

138. In July 2018, OpenAI wrote a document titled "Motivation," to which Alec Radford and Daniela Amodei contributed comments. Nelson Decl. Ex. 303.	138. **Undisputed.** Microsoft defers to OpenAI.
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.*	139. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 303 contains the quoted text. However, disputed that Books Ex. 303 establishes that LibGen specifically was identified as a source that could help OpenAI "explore how far this approach can go"; Books Ex. 303 does not make that connection and is from 2018, two years prior to GPT-3 training. Further disputed that Books Ex. 303 accurately frames OpenAI's discussion of LibGen, as its only comment on using LibGen is that it would require "significant work on collecting [and] filtering." Books Ex. 303 at OPCO_MDL_001938646. Microsoft defers to OpenAI.
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	140. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 243 contains the quoted text. However, disputed that Plaintiffs' characterization of,

that recent books were under-represented. *Id.* Radford responded that "modern books are very under-represented due to copyright." *Id.* at -899.	and inferences drawn from, Books Ex. 243 accurately reflect the context of the quoted statements. Radford asked for "links to high-quality/important text/articles they think everyone should read" to "stress test a new web-crawler's coverage." In response, Brundage stated that "recent books (i.e. non-Gutenberg) might be underrepresented," and Radford responded that "modern books are very under-represented due to copyright and what not." Disputed that these statements establish that modern books were underrepresented in OpenAI's training data generally or reflect OpenAI's training-data needs because the statements concerned the web crawler's coverage, including of a dataset consisting of "every url ever submitted to reddit" with karma filtering. Books Ex. 243 at OPCO_MDL_000220898–899. Microsoft defers to OpenAI.
141. In October 2018, OpenAI received an offer to buy "[redacted]." Nelson Decl. Ex. 332 at -638. Dario Amodei responded that "[redacted]" *Id.* at -637.	141. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 332 contains the quoted text. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Books Ex. 332 accurately reflect Amodei's assessment of the quality and potential use of the proposed dataset because they omit qualifying context. Amodei further stated, "[redacted]" Books Ex. 332 at OPCO_MDL_004585637. Microsoft defers to OpenAI.

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.*	142. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 332 contains the quoted text. However, disputed that Plaintiffs' characterization of Radford's response as expressing "disagreement" with Amodei accurately reflects the cited exchange, or that Radford's statements establish he disagreed with Amodei's assessment rather than identifying logistical considerations to evaluate before proceeding. Amodei asked Radford, "what are your thoughts? Do you think it's worth talking to them?," and Radford responded by identifying logistical considerations concerning the feasibility and potential scope of the proposed digitization project. Radford further stated, "[a] quick sanity check on that kind of logistics stuff would be important before proceeding." Books Ex. 332 at OPCO_MDL_004585637. Microsoft defers to OpenAI.
143. A few days later, Daniela Amodei followed up in the thread: "Coming back to this – [redacted] That works for me if so[.]" *Id.*	143. **Undisputed.** Microsoft defers to OpenAI.
144. Radford responded: "That's my current leaning, yep!" *Id.*	144. **Undisputed.** Microsoft defers to OpenAI.
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.	145. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: though the cited testimony only concerns [redacted] it is undisputed that OpenAI did not

	purchase books in 2018. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that testimony suggest that a functioning market for licensing trade books, including the Asserted Works, for LLM training existed in 2018. Radford's testimony does not establish either proposition. *See also* G. Leonard Opp. Decl. Ex. A ¶¶ 287–291 (finding no established market in which Plaintiffs could participate to license their Asserted Works for LLM training). Microsoft defers to OpenAI.
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.	146. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 146 are disputed because they are not supported by the cited evidence. Specifically, Mr. Radford testified [redacted] *See* C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 81:11–82:16; ECF 1767-1 (Choffnes Decl. Ex. A) ¶ 103. Microsoft defers to OpenAI.
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.	147. **Undisputed.** Microsoft defers to OpenAI.

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.	148. **Undisputed.** Microsoft defers to OpenAI.
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.	149. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the script downloaded approximately 110,000–120,000 files (not books). *See* C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 24:12-18. However, disputed that Plaintiffs' cited materials support the assertion that "the script ran for approximately one month," because the cited materials do not establish that length of time. Microsoft defers to OpenAI.
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.	150. **Undisputed.** Microsoft defers to OpenAI.
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.	151. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 151 are disputed because they are not supported by the cited evidence. Plaintiffs' cited materials do not support the assertion that Radford "wrote a script that would have restricted the download to books listed for sale on Amazon.com" or that "he did not run that script." When asked [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 100:4–18. Further, Ex. 738 contains no support for the proposition that Radford used the purported Amazon dataset referenced in this fact. *See* Books Ex. 738, at Radford00000425. Microsoft defers to OpenAI.
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.	152. **Disputed – Not Supported by Admissible Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 152 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Plaintiffs' cited evidence does not establish that, "[a]t the time of his download, Radford knew LibGen was 'a Russian website' containing free books." Amodei testified [redacted]" (C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 113:1-12) which is hearsay not within an exception and incapable of being offered in an admissible form at trial. Fed. R. Evid. 802. Further, the cited testimony does not support Plaintiffs' characterization that LibGen contained "free books," because that phrase does not appear in the cited testimony. Undisputed that Radford separately testified [redacted]

	[redacted] See C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 113:14–15. Microsoft defers to OpenAI.
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.	153. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 45 contains the quoted testimony. However, disputed that Radford testified that he could have purchased all of the books downloaded from LibGen, as that is not reflected in the cited evidence. Further, disputed that Plaintiffs' characterization of, and inferences drawn from, Radford's testimony accurately reflect its context because Plaintiffs' attorney asked, [redacted] See C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 113:14–15; 148:6–8. Microsoft defers to OpenAI.
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.	154. **Undisputed.** Microsoft defers to OpenAI.
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.	155. **Undisputed.** Microsoft defers to OpenAI.

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.	156. **Undisputed.** Microsoft defers to OpenAI.
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.	157. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 247 contains the quoted statement attributed to Tom Brown. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Brown's statement establish that the LibGen data was restricted to books listed for sale on Amazon. Brown qualified his statement with "I believe," and Radford stated in the same discussion that [redacted] Books Ex. 247 at OPCO_MDL_000478436. Further, the disputed proposition is not supported by the cited evidence. Further, the record does not establish that an Amazon-based filter was applied to the LibGen data. Radford testified that he [redacted] C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 100:4–18. Additionally, OpenAI expert Frederiksen-Cross's analysis of the resulting dataset demonstrates that the Amazon data was not used to filter the LibGen dataset. Because the Amazon product dataset includes no works after July 2014, yet Plaintiffs themselves identify multiple asserted works in the LibGen training datasets that were published after that date, the data is inconsistent with the assertion that an Amazon-based filter

	was applied. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 159–166. Microsoft defers to OpenAI.
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.*	158. **Disputed in Part.** OpenAI responds as follows: Undisputed that Ex. 283 contains the quoted text. However, disputed that the LibGen dataset was in fact filtered to books listed for sale on Amazon. Mr. Radford testified that, although he [redacted] C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 100:4–18. OpenAI expert Frederiksen-Cross's analysis of the resulting dataset confirmed that Radford was incorrect and that the Amazon data was not used to filter the LibGen dataset. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 159–166. Microsoft defers to OpenAI.
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.	159. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Radford's files were produced at Bates OPCO_SDNY_1736878 through -6901 and that Shan asserts that he identified 96 of Plaintiffs' Works in those files. However, disputed that the cited evidence shows that "96 of Plaintiffs' Works" are in fact located in those files. Specifically, Shan's identification methodology is flawed and produces false positives, undermining the reliability of the claim that he identified "96 of Plaintiffs' Works" in the

	referenced file set. S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 41–42, 72–77. Microsoft defers to OpenAI.
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.*	160. **Undisputed.** Microsoft defers to OpenAI.

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.	161. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the cited document contains the quoted text. However, disputed that "OpenAI revisited whether it should lawfully acquire books," because "lawfully acquire books" is an improper attorney argument and/or legal conclusion. It thus requires no response. Specifically, Plaintiffs' characterization that OpenAI's download from LibGen was unlawful, rather than fair use, is an improper legal conclusion. Microsoft defers to OpenAI.
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	162. **Undisputed.** Microsoft defers to OpenAI.

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.*	163. **Undisputed.** Microsoft defers to OpenAI.
164. Lightcap replied: "[redacted]" Nelson Decl. Ex. 345.	164. **Undisputed.** Microsoft defers to OpenAI.
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.*	165. **Undisputed.** Microsoft defers to OpenAI.
166. Lightcap responded: "[redacted]." *Id.*	166. **Undisputed.** Microsoft defers to OpenAI.
167. OpenAI did not move forward with buying books in 2019. Nelson Decl. Ex. 8 at 87:25–88:3.	167. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: though the cited testimony only concerns the purchase of books from [redacted], it is undisputed that OpenAI did not purchase books in 2019. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that testimony suggests that a functioning market for the use of

	trade books for LLM training existed in 2019. *See also* G. Leonard Opp. Decl. Ex. A ¶¶ 287–291 (opining that no established market in which Plaintiffs could participate to license their Asserted Works exists for LLM training). Microsoft defers to OpenAI.
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.	168. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 168 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the cited evidence is an email not written by Murati that purports to summarize what she said, and therefore constitutes inadmissible hearsay. Microsoft defers to OpenAI.
169. Murati suggested partnering with "[redacted]" *Id.* Murati noted that she had already "[redacted]" *Id.* Murati added that OpenAI could also "[redacted]" *Id.*	169. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 169 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the cited evidence is an email not written by Murati that purports to summarize what she said, and therefore constitutes inadmissible hearsay. Microsoft defers to OpenAI.

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.	170. **Disputed in Part – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Lansky participated in the email thread referenced in Books Ex. 293. Books Ex. 293 at OPCO_MDL_001839987. However, disputed that Lansky suggested that OpenAI acquire "free book downloads [from] libgen.io" or that libgen.io had "an API and ways to do bulk downloads." *See* C. Blythe Opp. Decl. Ex. 11 (D. Lansky 30(b)(1) Tr.) at 15:3-16:17, 32:17-33:2. Further, the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the cited evidence as to the disputed proposition is an email not written by Lansky that purports to summarize what he said, and therefore constitutes inadmissible hearsay. Microsoft defers to OpenAI.
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.	171. **Undisputed.** Microsoft defers to OpenAI.
172. Lansky testified that "I don't recall anybody objecting to my suggestion that they look at Library	172. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that

Genesis," and that he did not know whether OpenAI had any policy prohibiting the use of Library Genesis. *Id.* at 32:12–33:20.	Mr. Lansky so testified. However, disputed to the extent Plaintiffs contend or imply that OpenAI's use of materials from Library Genesis was unlawful, because any such use was a fair use. Accordingly, the Court should disregard the disputed portion of Paragraph 172 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Microsoft defers to OpenAI.
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.	173. **Undisputed.** Microsoft defers to OpenAI.
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.	174. **Undisputed.** Microsoft defers to OpenAI.
175. Lansky followed up that he had "some traction with Amazon / Kindle" and asked whether other employees were available to meet with them. *Id.* at -987.	175. **Undisputed.** Microsoft defers to OpenAI.
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.	176. **Undisputed.** Microsoft defers to OpenAI.
177. OpenAI [redacted]. On February 25, 2019, OpenAI employee Ben Chess copied the raw text out	177. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that OpenAI [redacted]

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.	[redacted]. Undisputed that Books Ex. 234 contains the quoted statement. Books Ex. 234 at OPCO_MDL_000136960. However, disputed that the cited evidence supports the assertion that "Ben Chess copied the raw text out of the files Radford had downloaded and created a dataset originally called 'libgen-reversible.'" Books Ex. 234 at OPCO_MDL_000136957. First, Chess testified [redacted] and reiterated "I don't -- I don't remember doing that." C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) at 81:5-10, 82:6-83:22. Second, Books Ex. 234—which S. Shan Decl. Ex. A ¶ 50 cites to support the same assertion—does not state that Chess "copied the raw text" of any data. Microsoft defers to OpenAI.
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.	178. **Undisputed.** Microsoft defers to OpenAI.
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.	179. **Disputed – Not Supported by Admissible Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 179 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Disputed that Books Ex. 300 establishes that "OpenAI described" the composition of LibGen1 as stated.

	Plaintiffs have not established that Ex. 300 was an official OpenAI memorandum or otherwise authenticated the cited statements as reflecting OpenAI's description of the dataset. The cited portion was not addressed during Chess's deposition, and when the same document was presented to Mann, [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 89:8-16. Further disputed because Books Ex. 300 contains inaccuracies. Radford testified that [redacted]. C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 48:1–4, 49:6–22. OpenAI further objects to Plaintiffs' reliance on the statements in Ex. 300 as unauthenticated hearsay and for lack of authentication. Fed. R. Evid. 802, 901; Fed. R. Civ. P. 56(c)(2). Microsoft defers to OpenAI.
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.	180. **Undisputed.** Microsoft defers to OpenAI.

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.	181. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 181 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, OpenAI objects to Books Ex. 13 because, to the extent this fact relies on Chess's testimony for the truth of what Radford purportedly proposed, it contains hearsay not within an exception and that is incapable of being offered in an admissible form at trial. Fed. R. Evid. 802. Books Ex. 13 therefore does not establish that Radford proposed that OpenAI use approximately 120,000 books downloaded from LibGen to train its models. Further disputed to the extent Plaintiffs' characterization suggests that the LibGen-1 dataset was proposed for, or used to train, *all* models at issue in this litigation. Microsoft defers to OpenAI.

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

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.	182. **Undisputed.** Microsoft defers to OpenAI.
183. OpenAI's co-founder and CEO Sam Altman then secured a meeting with Microsoft's CEO Satya Nadella,	183. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants do not dispute that

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.	OpenAI's co-founder and CEO Sam Altman secured a meeting with Microsoft's CEO Satya Nadella, but Defendants dispute the remainder of Paragraph 183. As to the second portion ("met with Nadella that very month"), Defendants dispute any implication that the timing of the meeting reflected urgency associated with the GPT-2 paper. Defendants also dispute that the cited evidence establishes that the timing of the meeting was driven by any such urgency. Defendants also dispute the characterization that "Nadella made clear that Microsoft was interested in technology transfer and commercialization with OpenAI." Mr. Altman testified that "commercialization" was "very undefined" at that point in time, Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 29:11-17, and Books Ex. 362 at OPCO_MDL_004851635 does not state that "Microsoft was interested in technology transfer." Microsoft also disputes any implication that Bill Gates was a Microsoft executive with decision-making authority over Microsoft's decision to invest in OpenAI. Mr. Gates did not have authority over Microsoft's decision to invest in OpenAI. *See* C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 109:3-110:2. Finally, Microsoft also responds the facts set forth in Paragraph 183 are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the cited evidence is hearsay and lacks foundation; the cited evidence are statements by

	OpenAI employees as to what Mr. Nadella was "interested in."
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.	184. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 317 contains the quoted text. However, disputed that Plaintiffs' characterization of Lansky's message as "follow[ing] up again about downloading LibGen" accurately reflects the cited evidence. Lansky asked whether anyone had "[dug] in to the Library Genesis link" he previously identified and noted that the link provided an option to download a database dump. Disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Lansky's statement suggest that he proposed obtaining books from LibGen for an unlawful purpose or that doing so was unlawful. *See* Books Ex. 317 at OPCO_MDL_002141166. Microsoft defers to OpenAI.
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.	185. **Undisputed.** Microsoft defers to OpenAI.
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.	186. **Undisputed.** Microsoft defers to OpenAI.

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.	187. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that in April 2019, OpenAI prepared a document for a meeting with Bill Gates that was also remotely attended by Umesh Madan, a technical advisor to Microsoft's CTO. But Microsoft disputes the claims that Bill Gates was a "Microsoft executive" in April 2019 and that "other Microsoft executives" attended the April 2019 meeting because this is not supported by the cited evidence. Bill Gates was no longer a Microsoft executive in April 2019 and did not have decision-making authority over Microsoft's investment in OpenAI. *See* C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 65:6-67:13, 109:3-110:2. Umesh Madan joined the in-person meeting by telephone to take notes, but no other Microsoft personnel were in attendance at the referenced April 2019 meeting. *See* Books Ex. 188 at -958; C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 15:24-16:1.
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.	188. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that Exhibit 812 states that OpenAI "added another ~11B words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context." But Defendants dispute the claim in Paragraph 188 that the document discussed in Paragraphs 187-188 described "the early results of OpenAI's training of the model eventually launched as GPT-3" or that the model or

	research described in Books Ex. 812 was part of the development of GPT-3 because the cited evidence does not support the stated fact. Exhibit 812 describes preliminary 2019 research into training models larger than GPT-2. Books Ex. 812 at OPCO_MDL_002135908. Books Ex. 812 does not mention GPT-3. Additionally, Plaintiffs' characterization is not consistent with the timing of GPT-3's development because OpenAI did not train GPT-3 until 2020. Books Ex. 316. Microsoft further responds that the facts set forth in Paragraph 188 are immaterial because these facts do not affect the outcome of the suit under the governing law. Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. Additionally, none of the testimony from Microsoft witnesses regarding the document cited as Nelson Decl. Ex. 812 demonstrates that Microsoft witnesses had personal knowledge of the sources of data in or contents of LibGen. C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:7-22; 52:20-25; C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 115:6-116:14, 119:2-14, 123:12-124:9.
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.	189. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that Exhibit 330 contains a draft presentation referring to "additional data sets, Book Corpus, LibGen, Wikipedia" and to a "6B and 12B parameter model." However, Defendants dispute that the presentation refers to "model[s]" in the plural, as

	Exhibit 330 refers to a singular model. Books Ex. 330 at OPCO_MDL_004584570. Defendants also dispute Paragraph 189 to the extent Plaintiffs' characterization of, and inferences drawn from, Books Ex. 330 suggest that the referenced model or datasets describe GPT-3 or the training data used to train GPT-3 because the cited evidence does not establish that connection. *Id.* Defendants also dispute that the cited testimony provides additional support for those assertions. Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 226:24–227:5. Defendants further dispute any implication in Paragraph 189 that the "draft OpenAI–Microsoft presentation" cited as Books Ex. 330 was the final version presented at the April 2019 meeting. Other evidence demonstrates that the presentation shown during the April 2019 meeting did not include a slide describing OpenAI's training datasets, and likewise did not include the language quoted in Paragraph 189. *See generally* C. Blythe Opp. Decl. Ex. 16; C. Blythe Opp. Decl. Ex. 17.
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.	190. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants do not dispute that Books Exs. 812 and 327 include a link to a Wikipedia webpage describing LibGen. But Defendants dispute the second sentence of Paragraph 190 because it is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the evidence cited (Books Ex. 670) is hearsay

	and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Microsoft further responds as follows: Microsoft disputes any implication in Paragraph 190 that the witnesses who received the document cited as Nelson Decl. Ex. 812 in connection with the April 2019 meeting followed the hyperlink to the Wikipedia page describing LibGen. *See* C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:17-22; C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 118:19-121:16.
191. Following the meeting with Gates, Microsoft agreed to invest $1 billion in OpenAI. Nelson Decl. Ex. 145 at -347; Nelson Decl. Ex. 203.	191. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** Defendants do not dispute the fact stated in Paragraph 191. But Defendants dispute any implication in Paragraph 191 of a causal relationship between the April 2019 meeting with Bill Gates and Microsoft's investment in OpenAI because this is not supported by the cited evidence. Books Exs. 145 and 203 do not reference the April 2019 meeting with Bill Gates. Indeed, Gates testified that he "advised Satya against doing the deal" and stated, "if I was in charge, I would go the internal route." C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 109:10-13. Gates further testified that "Satya made the choice to do a deal

	with OpenAI, which I, without trying to be too tough about it, said was not a deal that I would do." *Id.* at 94:10-12. Defendants further incorporate their responses to Paragraphs 187-190. Microsoft further disputes any implication in Paragraph 191 that the $1 billion investment was a cash investment. The 2019 commitment required minimum Azure spend of [redacted] *See* Books Ex. 147 at MSFT_AI_MDL_000000735. Further, Books Ex. 203 is a draft Microsoft memo that does not identify the author or recipient ("To:" and "From:" are blank) and therefore cannot be substantiated as containing correct information and is thus not admissible evidence, as required by Local Rule 56.1.

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.	192. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 192 are disputed because they are not supported by the cited evidence. Specifically, Brockman's cited testimony in Books Ex. 8 concerns [redacted], not whether OpenAI purchased or licensed books at any time after Microsoft's investment. See C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 87:25–88:3. Further disputed that "OpenAI did not purchase or license books" at any point after 2019. See

	OpenAI's response to Plaintiffs' facts 292 and 814. However, as noted in response to Plaintiffs' fact 814, [redacted], rather than trade books such as the Asserted Works and does not establish the existence or value of comparable transactions for Plaintiffs' works. *See also* G. Leonard Opp. Decl. Ex. A ¶¶ 293–298. Neither does [redacted] make such a showing. *See* Books Ex. 698. Microsoft defers to OpenAI.
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.	193. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 193 are disputed because they are not supported by the record. Specifically, the cited evidence does not establish that Mann and Hallacy *successfully* torrented approximately 35 terabytes of data from LibGen, because that calculation assumes completion of the torrent payloads, and the record does not support that assumption. OpenAI's expert Frederiksen-Cross's examination of the Hallacy LibGen Test File found that approximately 20% of files from the torrent payloads were incomplete or missing. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 141-144; Books Ex. 778 at OPCO_MDL_000498525; Books Ex. 801 at OPCO_MDL_000492368. Further, Mann testified [redacted]" C. Blythe Opp. Decl. Ex.

	13 (B. Mann 30(b)(1) Tr.) at 332:17-333:1; *see also* B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 144. Microsoft defers to OpenAI.
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.	194. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 194 are disputed because they are not supported by the cited evidence. Specifically, these facts are disputed for the reasons stated in response to Fact No. 193. Further, the cited evidence does not establish that the approximately 35-terabyte figure equaled the entirety of LibGen's fiction and nonfiction book collections because the cited evidence does not establish the volume of LibGen's fiction collection at the relevant time. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 146. Additionally, to the extent Plaintiffs' characterization of "[t]hat volume" implies that the LibGen fiction and nonfiction datasets were non-duplicative or consisted exclusively of English-language copyrighted books, those propositions are not supported by the datasets on which Shan relies. *See id.* Microsoft defers to OpenAI.
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.*	195. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Hallacy participated in the Data-thon. However, disputed that the cited evidence establishes that Mann sent the calendar invitation to Amodei or Brown because, although their names appear in the body of the

	invitation, Books Ex. 249 reflects the invitation being sent from Mann to Hallacy. Further disputed that Books Ex. 249 establishes that Mann participated in the Data-thon. While Books Ex. 249 establishes that Mann created the Google Calendar event and sent an invitation to Hallacy; it does not establish that Mann actually participated. Microsoft defers to OpenAI.
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.*	196. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited Google Calendar event contains the quoted language. However, disputed that Plaintiffs' characterization that the invitation "directed" participants generally to "[c]ollect" the specific referenced datasets accurately reflects the cited document. Rather, the document describes the event as concerning the collection of "language eval datasets" generally and lists several datasets as points of reference. See Books Ex. 249 at OPCO_MDL_000478674. Microsoft defers to OpenAI.
197. Hallacy described a "Data-thon" as a "day-long effort . . . to acquire datasets." Nelson Decl. Ex. 25 at 105:11–14.	197. **Disputed in Part.** OpenAI responds as follows: Undisputed that Hallacy described the Data-thon as a "day-long effort by several members of the language team to acquire datasets for their project." See Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 105:12-14. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Hallacy's testimony accurately reflect its context to the extent Plaintiffs suggest that the Data-thon involved

	acquiring books or other data from LibGen. Hallacy testified that the only dataset he recalled acquiring during the Data-thon was "a dataset of Amazon customer reviews," and when asked whether he remembered "going to LibGen or Library Genesis," Hallacy responded, "No." *See id.* 105:20-106:2. Microsoft defers to OpenAI.
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.	198. **Undisputed.** Microsoft defers to OpenAI.
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.	199. **Disputed in Part.** OpenAI responds as follows: Undisputed that the email cited as Books Ex. 250 contains the quoted text. Undisputed that Hallacy testified that he had used the "[redacted]" handle. See Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 106:15-107:20. However, disputed that the cited evidence establishes Hallacy registered for or accessed the Miner's Hut forum. Indeed, Hallacy testified that he did not recall signing up for or looking at Miner's Hut. *See Id.* 108:24-109:7. Microsoft defers to OpenAI.
200. Later that day, Mann wrote in an OpenAI Slack channel that "[H]allacy and [I] have begun downloading libgen." Nelson Decl. Ex. 260.	200. **Undisputed.** Microsoft defers to OpenAI.
201. Also on September 20, 2019, in another OpenAI Slack channel, Hallacy asked for help with "libgen	201. **Disputed in Part.** OpenAI responds as follows: Undisputed that in Books Ex. 308, Hallacy asked,

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.	"what's the easiest way to add a persistant disk to a dev box?" and undisputed that "disk" refers to storage. However, disputed to the extent Plaintiffs seek to draw any inferences or implications from the document. At deposition, Hallacy testified that he was "not sure what [libgen parsing] means in the context of this document," and when asked "[w]ith respect to LibGen, what does it mean to parse LibGen," Hallacy answered, "I have no idea." Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 113:22-114:16. Microsoft defers to OpenAI.
202. When Chess offered a solution, Hallacy responded, "easy enough. I shall do that." Nelson Decl. Ex. 308.	202. **Undisputed.** Microsoft defers to OpenAI.
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.	203. **Undisputed.** Microsoft defers to OpenAI.
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"	204. **Undisputed.** Microsoft defers to OpenAI.

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.	205. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited GitHub comment states, in full, "For libgen this helps to eliminate copyright info from start of document." However, disputed that Plaintiffs' characterization of, and inferences drawn from, that statement establish that OpenAI removed copyright information from LibGen documents. Specifically, OpenAI's expert Frederiksen-Cross explained that the referenced process randomly subsampled text from documents rather than always sampling the first 10,000 characters, and that this subsampling process does not remove any content from the underlying LibGen data, as "nothing is ever removed from the underlying data (or the good_df data frame) by this subsampling process." *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 154-58; C. Blythe Opp. Decl. Ex. 18 (*In Re OpenAI ChatGPT Litigation* ROG 18). Microsoft defers to OpenAI.
206. That Spark-based classifier removed copyright information from documents downloaded from LibGen. Nelson Decl. Ex. 25 at 119:21–120:9.	206. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 206 are disputed because they are not supported by the cited evidence. Specifically, Plaintiffs' characterization of, and inferences drawn from, Hallacy's testimony do not establish that the classifier was used to

	remove copyright information; Hallacy testified only that Spark "could" be used for that purpose. *See* Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 119:21-120:9. Further, no Spark-based classifier modified the underlying LibGen documents to remove copyright information from documents downloaded from LibGen. *See* C. Blythe Opp. Decl. Ex. 18 (*In Re OpenAI ChatGPT Litigation* ROG 18); B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 154-58. Microsoft defers to OpenAI.
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.	207. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that torrenting permits the transfer of data over a peer-to-peer network. However, disputed that Plaintiffs' characterization that participants "both download files from and upload files to one another" or that torrenting requires a participant to both download and upload files, including simultaneously, accurately describes torrenting or is supported by Mann's cited testimony. Rather, as OpenAI's expert Frederiksen-Cross explained, torrenting permits participants to download files from other participants on a peer-to-peer network without uploading. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 37, 39. Indeed, Plaintiffs' own expert, Choffnes, [redacted]. *See* C. Blythe Opp. Decl. Ex. 19 (D. Choffnes Tr.) at 70:12-19.

	Microsoft defers to OpenAI.
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.	208. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 802 contains the quoted statements. However, disputed that the cited evidence supports Plaintiffs' characterization that any "downloading" started on October 4, 2019. Mann testified [redacted]" C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 104:1-7. The "start downloading torrents" item on the October 4 to-do list bore a clock icon, not a check mark, indicating that Mann had not completed that task on October 4. *Id*. at 103:17-22, 104:8-11. Microsoft defers to OpenAI.
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.	209. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Transmission was the BitTorrent program Mann used, that the "0 upload" setting was not the program's default setting, and that Mann configured Transmission to disable uploading or seeding. *See* C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 413:11-416:18. However, disputed that the cited evidence supports Plaintiffs' characterization that Mann selected the "0 upload" setting "to prevent OpenAI from sharing books back to other users." The cited evidence — ECF 1767-1 (D. Choffnes Decl. Ex. A) ¶¶ 127, 129 — does not establish Mann's subjective reason for selecting that configuration.

	Choffnes's opinion regarding Mann's purpose in selecting the "0 upload" setting is speculation and unsupported by Mann's testimony or other record evidence. Microsoft defers to OpenAI.
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.	210. **Undisputed.** Microsoft defers to OpenAI.
211. On October 8, 2019, in the same channel, Mann estimated that torrenting "all [redacted]" 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.	211. **Undisputed.** Microsoft defers to OpenAI.
212. In the same OpenAI Slack channel, Mann posted a screenshot of the Transmission BitTorrent software managing the LibGen torrenting.  Nelson Decl. Ex. 268.	212. **Undisputed.** Microsoft defers to OpenAI.

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.	213. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 268 contains the quoted statement. However, disputed that Plaintiffs' characterization of, and inferences drawn from, the number of torrents queued establish that more than 2,000 torrents, or all files associated with those torrents, were successfully downloaded. Specifically, disputed that the number of torrents queued establishes the amount of data ultimately downloaded because, as OpenAI's expert Frederiksen-Cross explained, torrent yields can be materially below 100%. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 141-150. Microsoft defers to OpenAI.
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.	214. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the screenshot shows the IP address 35.232.155.37. Undisputed that OpenAI used Google Compute Engine to download data from LibGen. However, disputed that the Choffnes declaration is sufficient to support a factual claim that Google owns the IP address or that the IP address "reflect[s] that OpenAI used Google Compute Engine to torrent." Specifically, the Choffnes declaration cites no support for these statements. Microsoft defers to OpenAI.
215. The transfers ran through Google's IP address. Other users sharing the same files saw that address	215. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts

rather than an OpenAI address, which would have identified OpenAI as the party torrenting LibGen. *Id.*	set forth in Paragraph 215 are disputed in part. Undisputed that the referenced screenshot shows the IP address 35.232.155.37. However, the remaining propositions are disputed because they are not supported by the cited evidence. Specifically, Plaintiffs' characterization and inference about what other users saw or would have seen is disputed. Internet activity necessarily runs through service providers, so if OpenAI had not used Google for this network transfer, the relevant IP address would have belonged to a different network service provider, not "an OpenAI address." The Choffnes declaration cites no support to suggest otherwise. Microsoft defers to OpenAI.
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.	216. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the referenced document contain the quoted language. Disputed that Plaintiffs' characterization of a "rcall box" as an "internal computing environment" is supported by the cited evidence. The cited evidence states that "Mann figured out how to make it work, and said it was an issue with 'rcall,'" and does not characterize "rcall" as "internal computing environment." ECF 1767-1 (D. Choffnes Decl. Ex. A) ¶ 75. Microsoft defers to OpenAI.
217. Later on October 8, 2019, Mann told Brown that he was "pulling the rest of libgen" and had "all 2176	217. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows:

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.	Undisputed that Books Ex. 803 contains the quoted statements. However, disputed that each LibGen torrent file "contained approximately 1,000 books" because this claim is not supported by the cited evidence. Rather, the torrent files contained information to locate and potentially download up to approximately 1,000 files through the peer-to-peer network. Further disputed that Plaintiffs' characterization of, and inferences drawn from, the 2,176 queued torrents establish that Mann successfully downloaded more than 2.1 million books. Queueing a torrent does not establish successful completion of its payload; OpenAI's expert Frederiksen-Cross explained that queued or displayed torrents do not necessarily yield all associated files and found that approximately 20% of files in the Hallacy test torrent payloads were incomplete or missing. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 143-144. Microsoft defers to OpenAI.
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.*	218. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 329 contains the quoted text. However, disputed that Books Ex. 329 supports Plaintiffs' characterization of a pull request as a "proposed change to computer code" because this claim is not supported by the cited evidence. Microsoft defers to OpenAI.

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.	219. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 329 shows a folder called "torrents" within a directory called "libgen_test" in Hallacy's OpenAI Google Cloud Storage container. However, disputed that Plaintiffs' characterization of the folder as containing "the torrent files needed to download LibGen" is supported by the cited evidence or establishes that the folder contained torrent files sufficient to download LibGen. *See* Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 143:18-20. Microsoft defers to OpenAI.
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.	220. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 805 contains the quoted text. However, disputed that Plaintiffs' characterization of the 544 gigabytes as "books" accurately describes the data obtained from LibGen. Mann himself testified that, [redacted]." C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 261:21-22. Mann further testified that [redacted]. *Id.* at 147:4-14. Additionally, OpenAI's expert Frederiksen-Cross explained that the LibGen torrent payloads included duplicate, incomplete or missing files, and non-book content, including poems,

	catalogs, pamphlets, and academic papers. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 141-149. Microsoft defers to OpenAI.
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.	221. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 805 states "read alec's libgen code." Undisputed that Books Ex. 805 reflects that Mann posted a link titled "metadata-based libgen filter." However, disputed that Mann testified that "alec's libgen code" was "the prior art for how to download and make use of LibGen that [he] was leaning on." C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 140:17–141:20. Pursuant to Fed. R. Evid. 106, Mann actually testified that [redacted]" C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 140:17-141:20. Microsoft defers to OpenAI.
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.*	222. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 803 contains the quoted text attributed to Brown and Mann. However, disputed that Plaintiffs' characterization accurately reflects Books Ex. 803. The cited evidence reflects that Mann stated nonfiction was "10x bigger" than fiction, not that "all of LibGen" was "10x bigger." Further disputed that Brown "offered additional

	OpenAI resources to aid the torrent." The cited evidence reflects that Brown asked what he described as a "[p]otentially naive question" about whether it was possible to download in parallel. Brown did not offer additional OpenAI resources to aid the torrent. *See* Books Ex. 803 at OPCO_NDCAL_0868999. Microsoft defers to OpenAI.
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.	223. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Mann reported that the LibGen nonfiction collection contained approximately 1.4 million PDF files and roughly 500,000 other documents. Disputed that Plaintiffs' characterization of those approximately 2 million files as "books" is supported by the cited evidence. Books Ex. 373— which D. Choffnes Decl. Ex. A at ¶ 81 cites as support— does not state that the either the approximately 500,000 "other documents" or the "1.4 million PDF files" were books. Additionally, the number of files in the LibGen nonfiction collection is not necessarily equivalent to the number of unique books because, as OpenAI expert Frederiksen-Cross explained, the LibGen nonfiction collection included duplicate files, non-English entries, and non-book materials, including poems, catalogs, pamphlets, biographies, and academic papers. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 141-149. Further, disputed to the extent Plaintiffs contend that Mann's post that "libgen fiction is looking super high quality" conveyed any higher quality

	than not "totally garbled." C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 156:14-157:13. Microsoft defers to OpenAI.
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.*	224. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 687 reflects that Mann submitted a pull request titled "Dedup new libgen (#7)" and states, "Dataset lives at gs://ben-rcall/libgen/1021-clean/." *See* Books Ex. 687 at OPCO_MDL_001667050. Disputed that Plaintiffs' characterization of, and inferences drawn from, Exhibit 687 establish that the referenced Google Cloud Storage location contained "the dataset of books he had torrented," because the cited evidence does not describe the contents of that dataset. Microsoft defers to OpenAI.
225. On November 8, 2019, Dario Amodei committed a change to OpenAI's GitHub repository labeled "register libgen." Nelson Decl. Ex. 229.	225. **Undisputed.** Microsoft defers to OpenAI.
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	226. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 688 shows a pull request on November 18, 2019 and that Mann commented "Downloading now to my PD." *See* Books Ex. 688 at OPCO_MDL_001667053. Disputed that the cited evidence is sufficient to support the remaining factual allegations in

"new libgen," from which he took only mobi and epub file types. Nelson Decl. Ex. 301 at -624.	Paragraph 226. Specifically, the cited evidence does not support the contention that Mann "posted a link to 'new libgen.'" Books Ex. 688 at OPCO_MDL_001667053. And to the extent Plaintiffs use the documents to claim Radford received information, the documents contain hearsay not within an exception and that is incapable of being offered in an admissible form at trial. Fed. R. Evid. 802. Microsoft defers to OpenAI.
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.*	227. **Undisputed.** Microsoft defers to OpenAI.
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.	228. **Disputed.** OpenAI responds as follows: This fact is disputed because Plaintiffs' characterization of, and inferences drawn from, Books Ex. 257 do not accurately reflect the cited document. While Mann placed "[redacted], Plaintiffs omit that Amodei responded saying [redacted] See Books Ex. 257 at OPCO_MDL_000483758. Microsoft defers to OpenAI.

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.	229. **Undisputed.** Microsoft defers to OpenAI.
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.	230. **Undisputed.** Microsoft defers to OpenAI.
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.	231. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted text appears in Books Ex. 231. However, disputed that the cited evidence supports Plaintiffs' assertions. The exhibit indicates that someone submitted a pull request, but it is unclear from the exhibit itself that it was Mann who did so. Rather, the exhibit shows that Amodei asked a question in GitHub and Mann responded with the message in the body of the communication. Books Ex. 231 at OPCO_MDL_000135874. Further, Plaintiffs' assertion misrepresents the cited document. "Update libgen fiction and cc_giant, add bigger nonfiction (#1217)" was the title of the pull request, not a description of its purpose; the language "[b]ecause PDFs handle whitespace strangely and when you convert to text things like 2 column layouts or page footers get all jumbled up" does not describe the purpose of the pull request as Plaintiffs suggest. *Id.* The facts set forth in Paragraph 231 are therefore disputed as unsupported by the cited evidence. Microsoft defers to OpenAI.

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.*	232. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Exs. 25 and 801 contain the quoted statements attributed to Mann. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Mann's statement accurately reflect the stated focus of his Learning Day project because Plaintiffs omit qualifying context. Pursuant to Fed. R. Evid. 106, Mann stated that he planned to investigate PDF parsing techniques "esp those that work for math, such [as] mathpix, so that we can better extract our 32TB of libgen nonfiction." Books Ex. 801 at OPCO_NDCAL_0867245. The Mathpix examples referenced by Mann concerned parsing algebraic equations and other complex mathematical content. Books Ex. 801; Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 189:1-5. The cited evidence is thus not correctly represented. Microsoft defers to OpenAI.
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.	233. **Undisputed.** Microsoft defers to OpenAI.
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	234. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI produced billing logs reflecting ingress and egress volumes for several Google Compute

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.*	Engine "project[s]" during the relevant period, and that Dr. David Choffnes claimed to analyze those logs to estimate the volume of data transferred from the internet to the virtual machine used to run the LibGen torrents. However, disputed that Plaintiffs' characterization of, and inferences drawn from, those billing logs establish that the relevant project's data transfers were attributable solely to the LibGen torrents, because the relevant project was also used for other OpenAI initiatives. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 137-40. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
235. Mann confirmed that he "torrented both [the] fiction and nonfiction" components of LibGen. Nelson Decl. Ex. 33 at 145:2–11.	235. **Disputed in Part.** OpenAI responds as follows: Undisputed that Mann testified that he "torrented both fiction and nonfiction." However, OpenAI disputes Paragraph 235 to the extent Plaintiffs imply that Mann torrented all of LibGen's fiction and nonfiction collections, because his testimony does not identify the quantity of either collection. *See* C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 145:2–11. Microsoft defers to OpenAI.
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.	236. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the LibGen metadata cited by Plaintiffs in Books Ex. 262 identifies 2,182,642 files and 865,087 unique titles listed in LibGen's metadata. Disputed that Plaintiffs'

	characterization of those files as "book files" accurately reflects the cited evidence. The number of files identified in LibGen's metadata do not necessarily establish the number of books in the collection. OpenAI expert Frederiksen-Cross explained that files do not correspond one-to-one with books and include non-book content, such as poems and biographies. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 142–144. The disputed proposition is not supported by the cited evidence. Further disputed to the extent Plaintiffs' characterization suggests that these figures represent the number of files actually downloaded by OpenAI, because the cited analysis concerns LibGen metadata rather than the contents of OpenAI's download. B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 149. Microsoft defers to OpenAI.
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.	237. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Plaintiffs' cited evidence refers to approximately "32TB of libgen nonfiction." Disputed that Plaintiffs' characterization of the LibGen nonfiction collection as containing "more than 2 million books" is supported by the cited evidence. The number of files identified in LibGen's metadata does not necessarily establish the number of books in the collection. OpenAI's expert Frederiksen-Cross explained that files do not correspond one-to-one with books and include non-book content, such as poems and biographies. B. Frederiksen-

	Cross Opp. Decl. Ex. A ¶¶ 142–144. Further, disputed to the extent Plaintiffs' characterization suggests that these figures represent the number of files actually downloaded by OpenAI, because the cited analysis concerns LibGen metadata rather than the contents of OpenAI's download. *Id.* ¶ 149. Microsoft defers to OpenAI.
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.	238. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Mann processed approximately 452,790 files into the dataset known as libgen. Disputed that Mann drew files from the "nonfiction portion[]" of LibGen, which is not supported by the cited evidence. Disputed that Plaintiffs' characterization of those approximately 452,790 files as "books" is supported by the cited evidence. The number of files processed into LibGen2 does not necessarily establish the number of books; OpenAI's expert Frederiksen-Cross explained that files do not correspond one-to-one with books and include non-book content, such as poems and biographies. *See* B Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 142–144, 149. The disputed proposition is not supported by the cited evidence. Further, OpenAI disputes Plaintiffs' reliance on Nelson Decl. Ex. 300 because it contains objectionable hearsay and was not properly authenticated. Specifically,

	Books Ex. 300 was a version of the document marked Ex. 17 to the Chess deposition; however, the cited portion of this document [redacted]. The same document (previously marked Bates No. OPCO_NDCAL_1722276-1722279) [redacted]." C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 89:8-16. Radford [redacted]. *See, e.g.*, C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 48:1–4 [redacted]); *see also id.* at 49:6–21. Exhibit 300 was in no way an official OpenAI memo or business document, and Plaintiffs provide no additional evidence to corroborate their factual claim. Microsoft defers to OpenAI.
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.	239. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 300 recites the quoted language. Disputed that Plaintiffs may properly rely on Exhibit 300 because the cited evidence is incorrect and is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Books Ex. 300 was a version of

	the document marked as Exhibit 17 to the Chess deposition, but the cited portion of this document [redacted]. The same document (previously marked Bates No. OPCO_NDCAL_1722276-1722279) [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 89:8-16. Radford likewise [redacted] *See, e.g.*, C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 48:1-4; *see also id.* at 49:6-21. Books Ex. 300 was not an official OpenAI memo or business document, and Plaintiffs provide no additional evidence to corroborate their factual claim. Microsoft defers to OpenAI.
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.	240. **Undisputed.** Microsoft defers to OpenAI.
241. Across all of its LibGen downloads, OpenAI acquired millions of books. Shan Decl. Ex. A at ¶ 48.	241. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 241 are disputed because they are not supported by the cited evidence. Specifically, the Shan Declaration does not identify a methodology supporting Shan's calculation that OpenAI downloaded "some 4.83 million books." Further, Plaintiffs' characterization of

	downloaded files as "books" is not supported by the record. The number of files processed into LibGen2 is far less than "some 4.83 million" and does not necessarily establish the number of books; OpenAI's expert Frederiksen-Cross explained that files do not correspond one-to-one with books and include non-book content, such as poems and biographies. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 142–144, 149. Microsoft defers to OpenAI.
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.	242. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the Shan Declaration contains the stated language concerning the number of files included in the LibGen1 and LibGen2 training datasets. However, disputed that Plaintiffs' characterization of those files as approximately 570,000 "books" is supported by the cited evidence because LibGen files do not necessarily correspond one-to-one with unique books. B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 142. Further disputed that Shan's declaration establishes that OpenAI downloaded "millions of books" from LibGen or that "millions of the books" OpenAI purportedly downloaded were not used to train any OpenAI model; Shan's declaration does not support these assertions and does not identify the methodology supporting his calculation of the number of books purportedly

	downloaded. *Id.* ¶ 149. For those reasons, the disputed propositions are not supported by the cited evidence. Microsoft defers to OpenAI.
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.	243. **Undisputed.** Microsoft defers to OpenAI.

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]. at 113:11–14. McGrew added: [redacted] *Id.* at 113:15–18.	244. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 244 are disputed because they are unsupported by the cited evidence. Specifically, McGrew's cited testimony does not support Plaintiffs' characterization that OpenAI "maintained LibGen in centralized storage for years." McGrew did not testify to having personal knowledge of whether any LibGen datasets were stored in a centralized storage location, when they may have been first stored in a centralized storage location, or how long they may have been stored in a centralized location. McGrew testified only to his general understanding, "broadly speaking," that OpenAI [redacted]

	[redacted]. *See* Books Ex. 35 (R. McGrew 30(b)(1) Tr.) at 113:1-10. Further disputed that McGrew's testimony establishes that any researchers had access to, or obtained copies of, LibGen from OpenAI's centralized dataset storage for any purpose other than training. *See id.* at 113:15-18. Microsoft defers to OpenAI.
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.	245. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 245 are disputed because they are unsupported by the record. Disputed that Plaintiffs' cited evidence establishes that Hallacy downloaded approximately 86,984 files from LibGen in 2019 or that Hallacy downloaded or stored those files in his personal container in OpenAI's Azure storage account. C. Blythe Opp. Decl. Ex. 20 (M. Trinh 30(b)(6) Vol. 4 Tr. (Mar. 3, 2026)) at 21:20-23:24. Further disputed that Plaintiffs' characterization of the approximately 86,984 files as "books" is supported by the record; OpenAI expert Frederiksen-Cross explained that LibGen files do not correspond one-to-one with books and include non-book content, such as poems and biographies. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 142–144, 149. The assertion that Hallacy downloaded books "from LibGen in 2019" is further disputed for the reasons stated in response to Fact No. 193.

	Microsoft defers to OpenAI.
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.	246. **Disputed.** OpenAI responds as follows: This fact is disputed because Plaintiffs' characterization of, and inferences drawn from, Hallacy's testimony does not establish from his personal knowledge what data the identified Google Cloud containers contained or stored. Hallacy did not testify from personal knowledge that LibGen was stored in the [redacted]; he was interpreting information reflected in the spreadsheet presented to him. *See* Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 229:5-230:4. Additionally, Plaintiffs' references to "these Google Cloud containers" are vague and unintelligible, as they are not named in Fact No. 245 or any other Fact. Thus, any characterization of "these Google Cloud Containers" and their contents has not been connected, and is not applicable, to any other Fact. Microsoft defers to OpenAI.
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.	247. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Hallacy used the LibGen Test file to create unfiltered extracted LibGen samples that he shared in OpenAI Slack channels. However, disputed that Plaintiffs' cited evidence establishes that [redacted]. Trinh testified that he did not know whether OpenAI employees could access the LibGen Test file using [redacted]

	[redacted]. *See* C. Blythe Opp. Decl. Ex. 20 (M. Trinh 30(b)(6) Vol. 4 Tr. (Mar. 3, 2026)) at 63:1-64:20. Further disputed that the sharing of extracted samples in the [redacted] establishes that all OpenAI employees could access the underlying LibGen Test file, or even the samples, because the channel had a limited number of participants and there is no evidence that other OpenAI employees had access to Hallacy's personal container. *See* Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) 147:2-15; C. Blythe Opp. Decl. Ex. 20 (M. Trinh 30(b)(6) Vol. 4 Tr. (Mar. 3, 2026)) at 55:13-19; C. Blythe Opp. Decl. Ex. 21 (I. Sohl 30(b)(1) Tr. (Dec. 10, 2025)) at 54:5-15. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	248. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 248 are disputed because they are not supported by the record. Specifically, Plaintiffs' characterization of, and inferences drawn from, the naming convention do not establish that each folder actually contained 1,000 files or that each torrent necessarily yielded all 1,000 associated files. OpenAI expert Frederiksen-Cross explained that LibGen torrents had imperfect yields, such that files associated with a torrent could be incomplete or missing. B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 144.

	Microsoft defers to OpenAI.
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 "60b0dfd8939052146441 1d910d654ce7." Nelson Decl. Ex. 829 at 8.	249. **Undisputed.** Microsoft defers to OpenAI.
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.	250. **Undisputed.** Microsoft defers to OpenAI.
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.	251. **Undisputed.** Microsoft defers to OpenAI.

2. Who Could Access the Books

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.	252. **Disputed in Part.** OpenAI responds as follows: Undisputed as to the accuracy of the quote from Jeff Wu's testimony. However, disputed that Plaintiffs' characterization that Wu "sought out" LibGen-sourced materials accurately reflects his testimony. Pursuant to Fed. R. Evid. 106, [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 103:1-6. Microsoft defers to OpenAI.
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.	253. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited exhibits contain the quoted testimony. However, disputed that Plaintiffs' characterization of, and inferences drawn from, the cited testimony establish that LibGen data was broadly accessible within OpenAI or the scope of employees who actually had access to LibGen data. Pursuant to Fed. R. Evid. 106, Wu further testified [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 107:8-22; *See also* C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 145:21-25; C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 304:15-305:10; C. Blythe Opp. Decl. Ex. 21 (I. Sohl 30(b)(1) Tr. (Dec. 10, 2025)) at 47:15-48:13, 54:5-15, 77:5-24. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
254. OpenAI's data access policy provides that [redacted]	254. **Disputed in Part.** OpenAI responds as follows: Undisputed that [redacted]

[redacted] Nelson Decl. Ex. 361 at -377. OpenAI's dataset preparation code, [redacted] Shan Decl. Ex. A at ¶ 55.	references data stored in Yuan's personal container. However, disputed that Exhibit 361 provides that [redacted] Exhibit 361 [redacted] Books Ex. 361 at OPCO_MDL_004837377. The policy thus [redacted] Further disputed that Books Ex. 361 establishes the data-access policies applicable during the relevant period because the policy is dated June 2, 2023, after the LibGen data at issue had been deleted. *Id.*; *see also* C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 272:8-17. Microsoft defers to OpenAI.
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.	255. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 255 are disputed because they are unsupported by the cited evidence. Specifically, Shan's opinion that "a personal container is not private storage" and that storing data in a personal container left it "broadly accessible" is not supported by the record evidence. Yuan testified that the "personal" designation for storage accounts means they are [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 140:14-19. Further, Trinh testified that personal containers [redacted] C. Blythe Opp. Decl. Ex. 40 (M. Trinh 30(b)(6) Vol. 3 Tr. (Mar. 3, 2026)) at 53:9-54:2. Microsoft defers to OpenAI.
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.	256. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 43(a) contains the quoted testimony from Paino. However, disputed that the cited evidence supports Plaintiffs' characterization of Paino's testimony concerning OpenAI's access-control systems as testimony offered in his capacity as OpenAI's corporate designee. Paino was not designated under Rule 30(b)(6) on the topic of access controls, and accordingly his testimony on that subject cannot properly be attributed to OpenAI as a corporate admission. *See* Fed. R. Civ. P. 30(b)(6) (limiting designee testimony to noticed topics). Further disputed that Paino's testimony establishes the nature or scope of OpenAI's access controls. Paino himself testified that he was "not super familiar with the way that all our access-control systems work today," that he was merely "a user of them," and that access-control systems were "maintained by other teams." C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 305:15-17; *see also id.* at 304:15-305:10. The record evidence demonstrates that OpenAI did maintain access

	controls. C. Blythe Opp. Decl. Ex. 21 (I. Sohl 30(b)(1) Tr. (Dec. 10, 2025)) at 47:15-48:13 (explaining what access controls are and that [redacted]); *id.* at 54:5-15 (testifying that [redacted]); *id.* at 77:5-24 (testifying that the general practice of Sohl's team is [redacted]). Plaintiffs' characterization that OpenAI "did not have a robust access-control system" is therefore not supported by the cited evidence and is contradicted by the record. Microsoft defers to OpenAI.

3. Copies, Migration, and Retention

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" [redacted] files in "libgen1" at [redacted]/, and 518,769 files in "libgen2" at [redacted]. *Id.*	257. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that "libgen_overlap.ipynb" identifies three LibGen-derived datasets in Google Cloud Storage and records the stated file counts and storage locations for "original_libgen," "libgen1," and "libgen2." However, disputed that Plaintiffs' characterization of, and inferences drawn from, that the evidence establishes that OpenAI "made and retained multiple copies of the LibGen books." Specifically, the cited evidence—the "libgen_overlap.ipynb" file as described in the Choffnes Declaration—records the

	existence and file counts of three datasets in Google Cloud Storage but does not establish that OpenAI "made and retained multiple copies of the LibGen books" or that "libgen2" was a copy of "original_libgen" or "libgen1." For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
258. A second file, "libgen_dedup_join.ipynb," references at least five additional LibGen storage locations. Choffnes Decl. Ex. A at ¶ 134.	258. **Undisputed.** Microsoft defers to OpenAI.
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.	259. **Undisputed.** Microsoft defers to OpenAI.
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.	260. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 264 contains the quoted text. However, disputed that Plaintiffs' characterization of, and inferences drawn from, the phrase "all . . . raw libgen" establish that Mann migrated the entirety of LibGen's fiction and nonfiction collections to Azure. Specifically, disputed that Books Ex. 264 establishes that OpenAI acquired the entirety of LibGen's fiction and nonfiction collections. As OpenAI's expert Frederiksen-Cross

	explained, torrent payloads had incomplete or missing files and other limitations, the expected yield from a torrent file can be considerably less than 100%, and Mann himself recalled failed downloads. *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 142–49. Disputed that the reference to migration of "raw libgen" establishes that the underlying torrent data was preserved, because "raw libgen" was also used as a name for a dataset and the cited evidence does not link it directly to the underlying download data. *See* Books Ex. 264 at OPCO_MDL_000494633. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	261. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 261 are disputed because they are unsupported by the record. Specifically, the cited evidence shows, at most, that some LibGen-related datasets, including pre-filtered and post-filtered versions, existed on OpenAI's Azure blob storage at that time. However, Plaintiffs' characterization of those datasets as the "original" and "derived" versions of the data OpenAI downloaded is not supported by the cited evidence because the record does not establish any connection between those datasets and any specific download. Microsoft defers to OpenAI.

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.	262. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 262 are disputed because they are not supported by the record. Specifically, the cited evidence does not support Plaintiffs' characterization that OpenAI used LibGen data for "purposes beyond pretraining." [redacted] is part of the pretraining process—it is a necessary step in assembling the training dataset. The uses Plaintiffs identify—[redacted]—are all part of the pretraining process, not independent "purposes." *See* C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 243:9–246:21; *id*. at 247:1–6. Microsoft defers to OpenAI.
263. In September 2019, Dario Amodei "conclu[ded]" that "one of the best ways to access different versions of common [] crawl is to mix them in with" existing datasets like LibGen. Nelson Decl. Ex. 266 -at 255. Amodei directed Mann to "prioritize making it possible" to mix in common crawl with LibGen. *Id*.	263. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 263 are disputed because they are not supported by the record. Specifically, Books Ex. 266 does not contain the text Plaintiffs attribute to it. The document actually states "[redacted] not "[redacted]"

	[redacted] [redacted] [redacted]." *See* Books Ex. 266 at OPCO_MDL_000498255. Plaintiffs' quotation thus alters the original text in two material respects: it changes [redacted] to [redacted] and it substitutes [redacted] for [redacted]—changing the meaning from [redacted]. Further, Plaintiffs' characterization that Amodei directed Mann [redacted]" is not supported by the cited evidence. Amodei asked Mann [redacted]. *See id.* Microsoft defers to OpenAI.
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.	264. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 264 are disputed because they are not supported by the record. Specifically, Plaintiffs' characterization of LibGen as "the positive training class" is not supported by the cited evidence. The positive class was not limited to LibGen; rather, it consisted of a mix of WebText, Wikipedia, NY Times, and LibGen documents. *See* C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 244:4–10. Further, Plaintiffs' use of the undefined term "positive training class" inaccurately characterizes the classifier and its purpose. As OpenAI's corporate designee testified, the goal of the logistic regression classifier was to

	remove autogenerated SEO spam from Common Crawl by learning to distinguish between text believed to be human-authored (the positive class) and unfiltered Common Crawl data (the negative class), thereby slightly increasing the prevalence of human-authored text in the filtered Common Crawl dataset. *See id.* at 245:5–246:2. The cited evidence does not establish that LibGen alone served as the positive class against which OpenAI measured web-crawled data. Microsoft defers to OpenAI.
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.	265. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI stopped using LibGen to train its publicly released GPT models in late 2021. However, disputed to the extent Plaintiffs' generalized phrase "continued to use LibGen" suggests continued use of LibGen for pretraining or a broader or more significant post-2021 use than the cited evidence establishes. Plaintiffs' characterization is inaccurate and incomplete. Wu testified [redacted] *See* C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 115:2–116:1. For those reasons, the disputed proposition is not supported by the cited evidence.

	Microsoft defers to OpenAI.
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.*	266. **Disputed.** OpenAI responds as follows: This fact is disputed because the cited evidence does not contain the text Plaintiffs attribute to it. Books Ex. 306 states "[t]he big way to get a lot more data out of libgen is to solve PDF extraction," not "the best way to get more out of libgen is to solve PDF extraction." Plaintiffs' quotation thus materially alters the original text by substituting "best" for "big" and omitting the phrase "a lot more data," changing both the meaning and emphasis of Jain's observation. *See* Books Ex. 306 at OPCO_MDL_001939467. Microsoft defers to OpenAI.
267. In February 2022, Long Ouyang wrote that [redacted] because OpenAI has [redacted] that it was not then training on. Nelson Decl. Ex. 240 at -897. Those books remained in OpenAI's storage. *See id.*	267. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 267 are disputed because they are not supported by the record. Specifically, Plaintiffs' characterization of Ouyang's message as a declarative statement does not accurately reflect the cited evidence. Pursuant to Fed. R. Evid. 106, Ouyang did not write that [redacted] Rather, Ouyang asked a question: "my understanding is that [redacted] *See* Books Ex. 240 at OPCO_MDL_000208897. Plaintiffs' quotation thus alters the original text by substituting [redacted]

	for [redacted] and recasting Ouyang's tentative question as a definitive statement. Further, Plaintiffs omit Ouyang's qualification that the materials were not used for pretraining "because extraction is currently bad," which provides the actual reason for non-use. Additionally, disputed to the extent Plaintiffs imply the referenced [redacted] were sourced from LibGen, because the cited evidence does not identify LibGen as their source. *See id.* Also disputed that the cited evidence establishes that "[t]hose books remained in OpenAI's storage" because Books Ex. 240 does not support that assertion. Microsoft defers to OpenAI.
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.	268. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited evidence contains the quoted text. However, disputed that Plaintiffs' characterization that "OpenAI was still using the LibGen data at the time it deleted that data" accurately reflects the cited evidence or the context in which those statements were made. The cited messages from Wu appeared in the #project-clear Slack channel—a channel created to "excise Libgen from [OpenAI's] systems and storage"—while OpenAI was actively identifying LibGen-containing datasets for deletion, and Paino stated in the same channel that he was "compiling a list of paths to delete today." *See* Books Ex. 339 at OPCO_MDL_004608824_0001. Further

	disputed to the extent Plaintiffs' characterization suggests a use of LibGen independent of model training. Wu explained that [redacted] was being used to mix pretraining data into instruction-following fine-tuning, which is part of the training process, not an independent "use" of LibGen, and Wu noted that OpenAI could "continue just mixing in common crawl" as a substitute. *See* Books Ex. 339 at OPCO_MDL_004608824_0002. Microsoft defers to OpenAI.

G. April–May 2020: OpenAI Renamed LibGen "Books1" and "Books2"

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.	269. **Undisputed.** Microsoft defers to OpenAI.
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.	270. **Undisputed.** Microsoft defers to OpenAI.
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.	271. **Undisputed.** Microsoft defers to OpenAI.

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.	272. **Disputed in Part.** OpenAI responds as follows: Undisputed that OpenAI used datasets internally named "libgen1-dedup-clean" and "libgen1-dedup-clean-csam-filt" in training GPT-3 and GPT-3.5, respectively. However, disputed that Plaintiffs' characterization that these dataset names "reflected the same origin" establishes the nature or contents of the datasets or that the use of "libgen1" in an internal dataset name establishes that the dataset was identical to, or contained the same contents as, the original LibGen-sourced data. Further disputed to the extent Plaintiffs' characterization suggests that dataset names establish the relative significance of LibGen-derived data within the broader training-data mixtures used for GPT-3 or GPT-3.5. C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) at 49:9–24; Books Ex. 316. Microsoft defers to OpenAI.
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.	273. **Disputed in Part.** OpenAI responds as follows: Undisputed that OpenAI used datasets internally named "LibGen-2-dedup-clean," "libgen2-enonly-dedup-clean-csam-filt," and "libgen2-nonen-dedup" in training GPT-3 and GPT-3.5, respectively. However, disputed that Plaintiffs' characterization that these dataset names "reflected the same origin" establishes the nature or contents of the datasets or that the use of "libgen2" in an internal dataset name establishes that the dataset was identical to, or contained the same contents as, the original LibGen-sourced data. Further disputed to the extent

	Plaintiffs' characterization suggests that dataset names establish the relative significance of LibGen-derived data within the broader training-data mixtures used for GPT-3 or GPT-3.5. C. Blythe Opp. Decl. Ex. 21 (I. Sohl 30(b)(1) Tr. (Dec. 10, 2025)) at 36:13–15; C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) at 49:9–24; Books Ex. 316. Microsoft defers to OpenAI.
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.	274. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 811 contains Mann's quoted statements. However, disputed that Books Ex. 811 is "a draft of the paper that would announce GPT-3 publicly"; Books Ex. 811 identifies the document as the "200B Model Memo," an internal planning document, rather than "Language Models Are Few-Shot Learners," the paper that publicly announced GPT-3. *Id.* Disputed that Plaintiffs' characterization of Radford's "+1" reaction as recognizing "a good question or a point for discussion" is supported by the cited evidence or establishes an OpenAI-wide convention. C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 154:5-155:14. Further disputed that Plaintiffs accurately quote Subbiah's comment; Subbiah wrote, "what are the rules around including this dataset in research and publications?," not "using this dataset in research and publications." Books Ex. 811. Microsoft defers to OpenAI.

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.	275. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 228 contains the quoted text. However, disputed that Plaintiffs' characterization that Clark asked Mann "to add more specificity" accurately reflects the cited exchange; Clark posed a general editorial question, and Mann's response solicited further input. Books Ex. 228 at OPCO_MDL_000135666. Disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Mann's statement suggest that the description was deliberately vague because of copyright concerns. Mann separately testified [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 03, 2025)) at 409:7–16. Microsoft defers to OpenAI.
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.*	276. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 371 contains the quoted text. Disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Amodei's statement suggest that OpenAI described the datasets as Books1 and Books2 because of copyright concerns. Amodei's message raised a question concerning potential "FUD/reputational risks"; not that the naming decision was motivated by copyright concerns or an effort to conceal LibGen. Further disputed that Plaintiffs' characterization accurately reflects the

	broader context concerning naming Books1 and Books2. Radford [redacted] C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 210:13–211:8. Microsoft defers to OpenAI.
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.	277. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 254 contains the quoted statements. However, disputed that Brown was "OpenAI's research engineering lead for GPT-3." The disputed proposition is not supported by the cited evidence. Neither Exhibit establishes that Mr. Brown was in that role. *See* Books Ex. 254; Books Ex. 372. Further, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, the quoted exchange suggests that OpenAI's discussion of how to describe Books1 and Books2 reflected an effort to conceal the datasets because of copyright concerns. Plaintiffs' selective presentation of the exchange does not accurately reflect Radford's stated context for the discussion, including [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 10 **(**A. Radford 30(b)(1) Tr.) at 210:3–211:8. Microsoft defers to OpenAI.
278. Immediately before publication, Dario Amodei confirmed that OpenAI would "describe libgen as 'Books1' and 'Books2.'" Nelson Decl. Ex. 255 at -950.	278. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 255 contains the quoted language. However, disputed that Dario Amodei "confirmed" that OpenAI would describe LibGen as "Books1" and "Books2." The cited evidence does not reflect Amodei providing a confirmation; rather, Amodei was seeking confirmation from others, asking: "also confirming we describe libgen as 'Books1' and 'Books2' -- are we now ok with this cc @Ashley Pilipiszyn @Jack Clark?" *See* Books Ex. 255 at OPCO_MDL_000482950. Microsoft defers to OpenAI.
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.	279. **Disputed in Part.** OpenAI responds as follows: Undisputed that the published GPT-3 paper refers to "Books1" and "Books2" and does not identify those datasets as "LibGen" or "Library Genesis." However, disputed that Plaintiffs' characterization that OpenAI published the paper "without disclosing LibGen" accurately reflects the record evidence to the extent it implies that OpenAI had an obligation to identify LibGen by name, or that the company deliberately concealed LibGen. Mann testified that [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 78:7-15; 79:6-9. When asked [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 10 (A. Radford 30(b)(1) Tr.) at 212:6-7. Radford also [redacted]. *Id.* at 210:14–211:8. Microsoft defers to OpenAI.
280. OpenAI did not obfuscate Wikipedia or Common Crawl as sources in the same paper. Nelson Decl. Ex. 33 at 310:20–311:7.	280. **Disputed in Part.** OpenAI responds as follows: Undisputed that the published GPT-3 paper identifies "Wikipedia" and "Common Crawl" by name. However, disputed that Plaintiffs' characterization of, and inferences drawn from, that fact establish that OpenAI selectively "obfuscated" LibGen while identifying other training-data sources. Further disputed to the extent Plaintiffs' comparison of LibGen with Wikipedia and Common Crawl suggests that OpenAI's naming convention was selectively applied to LibGen. Mann testified that [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) at 79:6-9. The GPT-3 paper also uses the name "WebText2" for a non-standard dataset derived from Reddit-linked content rather than identifying it by its underlying source. Books Ex. 162 at MSFT _AI_MDL_000101661, Table 2.2. Microsoft defers to OpenAI.
281. OpenAI witnesses did not provide a non-privileged explanation for the renaming. Nelson Decl. Ex.	281. **Disputed.** OpenAI responds as follows: Disputed that OpenAI witnesses did not provide a non-

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.	privileged explanation for the renaming. OpenAI witness Ben Mann testified [redacted] Books Ex. 693 at 394. Microsoft defers to OpenAI.
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.	282. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited evidence contains the quoted testimony. However, disputed to the extent Plaintiffs suggest that Mann did not provide a non-privileged explanation for why the names "Books1" and "Books2" were used. Mann separately testified [redacted] Books Ex. 693 at 394. Further disputed that the testimony of Amodei, Ryder, Wu, and OpenAI's corporate designee reflecting a lack of recollection, involvement, or awareness of naming decisions establishes that OpenAI lacked a non-privileged explanation for the naming of Books1 and Books2. Microsoft defers to OpenAI.

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.	283. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the #large-book-dataset Slack channel was created. However, disputed that Plaintiffs' cited evidence establishes Altman's participation in or awareness of the channel because Altman is not identified as a member of the cited Slack thread in Books Ex. 694, and the Lasinski Declaration does not establish that Altman was alerted to the channel. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.	284. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 370 contains the quoted statement. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Jain's statement establish that Jain was proposing that OpenAI obtain additional material from LibGen or that "get more out of libgen than we have currently" refers to acquiring additional LibGen material. The cited message instead refers to obtaining more usable text from existing LibGen PDFs through improved [redacted]. Jain's LibGen comment in Books Ex. 370 was one part of a broader set of data-strategy updates that also addressed extracting new Common Crawl dumps, developing a distributed scraper, and improving HTML-to-text extraction for

	newspapers and forums. Books Ex. 370 at OPCO_NDCAL_0001195. Microsoft defers to OpenAI.
285. In 2021, Sutskever wrote that "[redacted]." Nelson Decl. Ex. 358. John Schulman responded that [redacted]. *Id.*	285. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 358 contains the quoted statements. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Sutskever's statement establish that OpenAI sought or planned to purchase [redacted]. Sutskever posed the broader question, "should we just buy all the books that exist?," as part of a strategic discussion concerning privacy considerations in the training process. Books Ex. 358 at OPCO_MDL_004792259. Further disputed that Plaintiffs' characterization of Schulman's response accurately reflects his statement. Schulman did not state that "the files came in an inconvenient format"; rather, Schulman commented that [redacted] "is very privacy averse and doesn't give you an epub file or anything." *Id.* Microsoft defers to OpenAI.
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.*	286. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 358 contains Schulman's quoted statement. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Schulman's statement establish that OpenAI adopted or implemented a strategy of purchasing books and separately downloading copies from LibGen. *See* Books Ex. 358 at

	OPCO_MDL_004792260. Further disputed to the extent Plaintiffs' selective quotation suggests Schulman concluded this would "cover" OpenAI legally; pursuant to Fed. R. Evid. 106, Schulman immediately qualified his suggestion by stating, "i don't know how the copyright laws work and whether this would cover us." *Id.* Microsoft defers to OpenAI.
287. In 2021, Radford wrote that [redacted]." Nelson Decl. Ex. 694. Radford estimated that [redacted] *Id.* at -095.	287. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 694 contains the quoted text. However, disputed that Plaintiffs' bracketed insertion [redacted] accurately reflects the cited text or that Plaintiffs' characterization of Radford's statements as proposing to [redacted] is supported by the cited evidence; Radford discussed [redacted] Books Ex. 694 at OPCO_MDL_004641094, OPCO_MDL_004641095. Further disputed because Radford wrote, [redacted] which Plaintiffs alter by replacing the question mark with a period. *Id.* at OPCO_MDL_004641095. Further disputed that Plaintiffs' characterization of, and inferences drawn from, Radford's statements accurately reflect the context of the discussion because Plaintiffs omit Radford's discussion bearing on his intent to extract book text for purposes other than

	distribution. *Id.* at OPCO_MDL_004641094, OPCO_MDL_004641095. Microsoft defers to OpenAI.
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."  *Id.*	288. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 694 contains the quoted statement. However, disputed that Radford's reference to avoiding "protracted discussions / negotiations / licensing" establishes an intent to avoid licensing requirements rather than a preference to avoid lengthy negotiations. Books Ex. 694 at OPCO_MDL_004641095. Pursuant to Fed. R. Evid. 106, Plaintiffs omit Radford's express qualification that he "understand[s] [he's] pretty naive on the other org decision factors involved." *Id.* Further disputed that Radford's stated preference establishes OpenAI's organizational position or strategic decision, particularly where Radford expressly qualified his views concerning other organizational considerations. *Id.* Microsoft defers to OpenAI.
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.*	289. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 694 contains the quoted text. However, disputed that Plaintiffs' characterization that Radford said "having contractors do it would be fine" accurately reflects the cited text because Plaintiffs omit the full quote; pursuant to Fed. R. Evid. 106, Radford discussed the cost and overhead of using contractors to make purchases: "I guess I could also imagine quickly doing

	numbers on what the contractor overhead would be ([redacted]). But yeah seems like it's possible to script this up on something like amazon." Books Ex. 694 at OPCO_MDL_004641095. Microsoft defers to OpenAI.
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.*	290. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 694 contains the quoted statements. However, disputed to the extent Plaintiffs' characterization suggests that OpenAI endorsed the proposed approach, because the cited exchange reflects a brainstorming discussion in which participants expressly raised licensing and compliance concerns and proposed consulting OpenAI's legal team regarding the appropriate approach. *See* Books Ex. 694 at OPCO_MDL_004641095. Plaintiffs' selective quoting omits Radford's qualification; pursuant to Fed. R. Evid. 106, Radford testified, "[m]aybe all of this is non-starter from licensing / compliance though?" and Brockman's response, "i think that from a licensing standpoint, we should definitely work with @David Lansky et al to find the best approach." *Id.* Microsoft defers to OpenAI.
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.	291. **Undisputed.** Microsoft defers to OpenAI.

292. [redacted]. Lasinski Decl. Ex. A. at ¶ 102. [redacted] Nelson Decl. Ex. 698. [redacted] *Id.*	292. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that Books Ex. 698 contains the quoted statements. *See* Books Ex. 698. However, disputed as to "at a scale sufficient to replace LibGen" because it is undefined, vague, and unclear. The Lasinski report does not reach this conclusion or make this characterization, and neither of Plaintiffs' cited sources specify what "replace" means and for what purpose. Microsoft defers to OpenAI.
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.*	293. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 321 contains the quoted statements. *See* Books Ex. 321. However, disputed that Plaintiffs' characterization of, and inference drawn from, Books Ex. 321 establish that OpenAI was evaluating the cost of purchasing books specifically "to replace those it had download from LibGen." The disputed proposition is not supported by the cited evidence. Specifically, Books Ex. 321 reflects a cost-scoping exercise in the context of OpenAI's broader data-planning strategy and attempts to quantify the cost in terms of token size as compared to the token sizes of other databases. Books Ex. 321 does not establish that the purpose of purchasing books was to replace previously downloaded LibGen data, rather than to expand or supplement OpenAI's training corpus. *See* Books Ex. 321.

	Microsoft defers to OpenAI.
294. Later in 2022, Brockman wrote to Sutskever that "books is like the best [redacted] 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.	294. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 347 contains the first two quoted statements. However, disputed that the exchange reflects any discussion of negotiations with book publishers. There is no mention of book publishers or negotiations with them in the referenced exhibit. Microsoft defers to OpenAI.
295. 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. [redacted]	295. **Undisputed.** Microsoft defers to OpenAI.
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.	296. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the testimony cited testimony contains the quoted language. However, disputed that Plaintiffs' characterization of Shetty's testimony as that of OpenAI's "corporate representative" is supported by the cited evidence. Shetty was not designated under Rule 30(b)(6) on

	the topic of purchasing or scanning books; his designated topics were limited to data-access agreements, publisher-partnership agreements, and Media Manager. C. Blythe Opp. Decl. Ex. 24 (V. Shetty 30(b)(6) Tr.) at 9:3–10:4. Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Shetty's lack of personal awareness establish that OpenAI did not purchase or scan books at scale. Shetty testified that he had never seen the [redacted], was not familiar with [redacted], and identified Michelle Fradin as the person with knowledge on the subject. C. Blythe Opp. Decl. Ex. 24 (V. Shetty 30(b)(6) Tr.) at 232:20–233:16. Moreover, OpenAI objected at the deposition on scope grounds before Shetty stated he was "not aware" of book purchasing en masse. *Id. at* 233:25–234:2. The Court should disregard Paragraph 296 to the extent the cited evidence does not support Plaintiffs' characterization. Microsoft defers to OpenAI.
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.*	297. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 67 contains the quoted testimony. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Yuan's testimony suggest that his statement concerning token-acquisition cost establishes the particular value of books to LLM performance. Yuan's statement that "you can reuse the data that you acquired" was not specific to books or e-books and applies more generally to acquired data. C.

	Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 159:9–14. Further disputed to the extent Plaintiffs suggest that Yuan's testimony establishes that a high-quality, diverse e-books dataset would improve LLM performance; the cited testimony does not establish that proposition. See *id.* Microsoft defers to OpenAI.
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.	298. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 695 contains the quoted text. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Books Ex. 695 establish that [redacted]; the cited evidence reflects OpenAI's consideration of [redacted] it. Further disputed that Lightcap observed that [redacted] "likes to push the limits of copyright." Books Ex. 695 reflects that Peter Hoeschele, not Lightcap, made that statement. Books Ex. 695 at OPCO_MDL_004620387_0003. The Court should disregard the disputed portions of Paragraph 298 because the cited evidence does not support those portions of the asserted fact. Microsoft defers to OpenAI.
299. OpenAI explored obtaining books from [redacted] but did not obtain any. Nelson Decl. Ex. 8	299. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 8 contains the quoted

at 176:8–177:19. Brockman confirmed that OpenAI [redacted] *Id.* at 124:14–18.	testimony. However, disputed to the extent Plaintiffs imply that Brockman confirmed from personal knowledge that OpenAI discussed obtaining [redacted]. Brockman did not testify to a personal recollection of such discussions. Rather, he testified: [redacted] and when asked whether that included books, stated: [redacted] C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 124:14–18. Brockman's testimony thus reflects his expectation, not his personal recollection or confirmation of the substance of any such discussions. Microsoft defers to OpenAI.
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.	300. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 9 contains the quoted text. However, disputed to the extent Plaintiffs' characterization of Brundage's testimony suggests that he had personal knowledge that a proposed book partnership was treated as low priority because OpenAI already had LibGen. Brundage [redacted]

	[redacted]. C. Blythe Opp. Decl. Ex. 32 (M. Brundage 30(b)(1) Tr.) at 290:6–24. Microsoft defers to OpenAI.
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*.	301. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 289 contains the quoted statement. However, disputed that Plaintiffs' characterization of, and inferences drawn from, this quote establish that OpenAI "determined" public-domain books could not substitute for LibGen books. The cited Slack exchange reflects an initial experiment, and subsequent record evidence identified OCR errors in [redacted] dataset as a principal limitation. C. Blythe Opp. Decl. Ex. 25 at OPCO_MDL_001940375 (Jan. 21, 2022 message from N. Ryder: "current plan is not to [include public domain books dataset Yuan tested on] too many OCR errors unfortunately"). Yuan testified that he did not observe performance degradation on the [redacted] "is no longer an eval that OpenAI cares about." C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 157:4–5; 168:10–11. Further disputed that OpenAI determined public-domain books were unsuitable for training, because record evidence reflects that OpenAI improved [redacted] for use. C. Blythe Opp. Decl. Ex. 26 at

	OPCO_MDL_001903117–18; C. Blythe Opp. Decl. Ex. 27 at OPCO_MDL_001898028; C. Blythe Opp. Decl. Ex. 28 at OPCO_MDL_000115341; C. Blythe Opp. Decl. Ex. 29 at OPCO_MDL_004631319. The Court should disregard the disputed portions of Paragraph 301 because those portions are not supported by the record. Microsoft defers to OpenAI.
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.	302. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 302 are disputed for the same reasons stated in OpenAI's response to Fact No. 301, which is incorporated by reference herein. Microsoft defers to OpenAI.
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.	303. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 303 are disputed because they are not supported by the record. Specifically, Plaintiffs' characterization that substituting public domain books "required more compute" is inaccurate. Yuan testified that he "[did not] see any evidence" that dropping LibGen would require OpenAI to spend more money on compute: "Q Did removing LibGen from OpenAI's training datasets require OpenAI to spend more money on compute? A I don't see any evidence that's suggesting that. And my personal opinion is it doesn't." C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1)

	Tr.) at 156:23–157:1. This fact is further disputed for the same reasons stated in OpenAI's response to Fact No. 301, which is incorporated by reference herein. Microsoft defers to OpenAI.
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.*	304. **Undisputed.** Microsoft defers to OpenAI.

I. December 2021–Summer 2022: OpenAI Deleted LibGen

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.	305. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 323 contains the quoted text. However, disputed to the extent Plaintiffs' characterization inaccurately suggests that Mr. Schulman had first-hand knowledge of Anthropic's model training practices and was seeking to override OpenAI's lawyers. The record evidence establishes that Mr. Schulman's quoted comment about Anthropic was [redacted] C. Blythe Opp. Decl. Ex. 30 (J. Schulman 30(b)(1) Tr.) at 140:21–141:4, 141:15–21.

	Microsoft defers to OpenAI.
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.	306. **Undisputed.** Microsoft defers to OpenAI.
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*.	307. **Undisputed.** Microsoft defers to OpenAI.
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*.	308. **Undisputed.** Microsoft defers to OpenAI.
309. Schulman responded at 4:25 p.m. that he was "not optimistic about getting the books from legal sources, and [redacted]." *Id*.	309. **Undisputed.** Microsoft defers to OpenAI.
310. OpenAI meeting notes from December 2021 state: "can't use libgen due to legal reasons." Nelson Decl. Ex. 320 at -476.	310. **Undisputed.** Microsoft defers to OpenAI.
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.	311. **Undisputed.** Microsoft defers to OpenAI.
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,	312. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted Slack messages were sent or that Paino added Weng, Wu, and Pachocki to the channel.

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.	However, disputed Plaintiffs' characterization that McGrew identified "GitHub storage" as a category of material to remove. Pursuant to Fed. R. Evid. 106, Books Ex. 338 states: "Github storage - do we have it in git?"—a question about whether the material was in GitHub. Books Ex. 338 at OPCO_MDL_004608820. The record further establishes that GitHub code was retained, and OpenAI subsequently recovered the Books1 and Books2 datasets used to train GPT-3. C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) at 69:4-18, 86:13–87:12. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.*	**313. Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 338 contains the quoted language. However, disputed that Plaintiffs' selective quotation of McGrew's message accurately reflects the context of the discussion concerning removal of LibGen-related materials. Plaintiffs omit McGrew's additional question. Pursuant to Fed. R. Evid. 106, McGrew asked: "Are there any upcoming experiments where we would like to repro GPT-3, or are we past that point?" Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, the exchange suggest that McGrew's direction to remove LibGen-related materials was unrelated to OpenAI's assessment of whether the data remained necessary for research or model-reproduction purposes;

	McGrew expressly raised whether OpenAI had any upcoming need to reproduce GPT-3 as part of the same discussion. Books Ex. 338 at OPCO_MDL_004608820_0002. Microsoft defers to OpenAI.
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.*	314. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 35 and Books Ex. 43(b) contain the quoted testimony. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, this testimony imply that Project Clear suggests prior acquisition or use of LibGen violated copyright law. Microsoft defers to OpenAI.
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.	315. **Undisputed.** Microsoft defers to OpenAI.
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.	316. **Undisputed.** Microsoft defers to OpenAI.

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.	317. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 338 contains the quoted testimony. However, disputed that Plaintiffs' characterization that McGrew "directed" OpenAI to remove such components by July 1 accurately reflects the cited evidence; McGrew stated that OpenAI "would like" to remove them by that date. Books Ex. 338 at OPCO_MDL_004608820. The record further establishes that GitHub code was retained, and OpenAI subsequently recovered the Books1 and Books2 datasets used to train GPT-3. C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) at 69:4-18. Microsoft defers to OpenAI.
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.	**318**. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 338 contains the quoted Slack exchange. However, disputed to the extent Plaintiffs' characterization of Weng as OpenAI's "then-Vice President of Research for Safety" suggests that she was responsible for decisions concerning training data; Weng testified that [redacted] C. Blythe Opp. Decl. Ex. 31 (L. Weng 30(b)(1) Tr.) at 14:13–22. Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, the exchange suggest that Weng proposed removing LibGen data that was actively being used for training; Weng clarified in her deposition [redacted]

	[redacted] *Id.* at 77:21–78:16. Microsoft defers to OpenAI.
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.	319. **Undisputed.** Microsoft defers to OpenAI.
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.	320. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Paino compiled a list of datasets. However, disputed that the list consisted of "datasets containing LibGen books that OpenAI had determined to delete." The cited evidence does not support that characterization. Books Ex. 222 shows Paino circulating a list of known locations of datasets and asking for confirmation on completeness. Books Ex. 222. Books Ex. 340 is a comprehensive inventory of all datasets and storage paths across OpenAI's systems, including Common Crawl, code datasets, Wikipedia, and many other non-LibGen datasets. The list was not limited to "datasets containing LibGen books," and characterizing the entire 18-page, 498-plus-entry list as such is misleading. Books Ex. 340. The Court should disregard the disputed proposition because the cited evidence does not support it.

	Microsoft defers to OpenAI.
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.	321. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 338 contains the quoted statement. However, disputed because the phrase "when OpenAI considered removing LibGen" is not supported by record evidence and is misleading, as the decision to remove LibGen datasets had already been made by June 2022, the date of Paino's comment in Exhibit 338. Books Ex. 43(b) (A. Paino 30(b)(1) Tr.) at 151:8-153:24. Further disputed to the extent Plaintiffs imply that Paino's comment in Exhibit 338 was part of the decision-making process to remove LibGen datasets. *Id.* Microsoft defers to OpenAI.
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.	322. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 338 contains the quoted statement. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Paino's statement suggest OpenAI was resistant to removing LibGen datasets or that Paino was reconsidering whether LibGen-related data should be removed. The decision to remove LibGen-related data had already been made when Paino made this statement. Paino raised the narrower question of whether to temporarily retain final filtered versions of LibGen1 and LibGen2 used for [redacted] until OpenAI determined that it was "OK with not being able to

	repro those results." Books Ex. 43(b) (A. Paino 30(b)(1) Tr.) at 155:13–156:4. Microsoft defers to OpenAI.
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.	323. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 43(a) contains the quoted Paino testimony. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that testimony suggest that LibGen data was still being used to pretrain OpenAI models at the time of deletion. Paino testified in the same exchange that OpenAI had already "stopped using LibGen to train -- pretrain models" before discussing the subsequent deletion of LibGen-derived data. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 272:2–13. Further disputed to the extent Plaintiffs imply that "Lib-Gen derived data" consists of datasets outside of those established by record evidence. Microsoft defers to OpenAI.
324. The LibGen datasets are the only training datasets OpenAI has ever deleted. Nelson Decl. Ex. 62(a) at 59:7–60:8.	324. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed that the cited testimony establishes that the LibGen datasets are the only training datasets OpenAI has ever deleted. Trinh testified only that he was "not aware of any" other datasets deleted outside of Books1 and Books2. Books Ex. 62(a) (M. Trinh 30(b)(6) Tr. (Jan. 29, 2025)) at 59:7–60:8. Trinh's lack of awareness does not establish and the

	disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.	325. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in this statement are disputed because they are not supported by the record. Specifically, disputed that the Choffnes Declaration establishes that "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." The cited paragraphs in Choffnes report are about the BitTorrent records, not the original set of files Mann downloaded. Nor does Choffnes attribute the loss of these files to the LibGen deletion. Disputed that torrenting requires a participant to both download and upload files, including simultaneously. Torrenting permits participants to download files from other participants on a peer-to-peer network without uploading. B. Frederiksen-Cross Opp. Decl. Ex. A ¶¶ 37, 39. Further disputed that OpenAI failed to produce "the versions of the LibGen datasets it used to train GPT-3.5." OpenAI produced the LibGen datasets used to train GPT-3 at bates numbers OPCO_SDNY_1457402 and OPCO_SDNY_1457403, and the LibGen data used to train GPT-3.5 was a slightly filtered subset of the GPT-3 data. See C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.)

	at 235:4-19; C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 421:14-422:8. Microsoft defers to OpenAI.
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.	326. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited evidence contains the quoted statements. However, disputed to the extent Plaintiffs imply that they could not analyze the contents of the GPT-3.5 dataset. OpenAI produced the LibGen datasets used to train GPT-3 and the LibGen Data used to train GPT-3.5 was a slightly filtered subset of the GPT-3 data. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 235:4-19. Plaintiffs' expert Shan acknowledges that the LibGen data used to train GPT-3.5 was a slightly filtered subset of the GPT-3 data and he states that "[a]s such, I understand that the books present in 'libgen1-dedup-clean' would also be present in 'libgen1-dedup-clean-csam-filt.'" ECF 1814 (S. Shan Decl. Ex. A) ¶ 51. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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	327. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI deleted LibGen-related data and that OpenAI witnesses testified that reproducibility can be important to research. However, disputed that Plaintiffs' cited evidence establishes that the deletion of LibGen data

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.	"prevented OpenAI from reproducing its earlier model training runs." Books Ex. 67 does not address LibGen; Yuan's cited testimony concerns reproducibility in the context of Common Crawl data, not LibGen. C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 23:18–24:5. Further disputed to the extent Plaintiffs suggest OpenAI witnesses testified that reproducibility of LibGen-based training runs was important to OpenAI's research. Yuan testified generally that reproducibility is important for research or scientific purposes, and Paino testified that reproducing prior research can be useful in certain circumstances; neither cited testimony establishes the asserted proposition specifically as to LibGen and without qualification. *Id.* at 23:18–24:5; Books Ex. 43(b) (A. Paino 30(b)(1) Tr.) at 152:6–153:24. For those reasons, the disputed propositions are not supported by the cited evidence. Microsoft defers to OpenAI.
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.	328. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI did not advance a "solely practical or technical reason for the deletion." However, disputed that Plaintiffs' cited evidence supports the remainder of Plaintiffs' statement, because the cited testimony does not discuss or support that characterization. Specifically, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that testimony suggest that LibGen data was still being used to pretrain OpenAI models at the

	time of deletion. The cited testimony does not discuss whether LibGen data was still being used for pretraining at the time of deletion. *See* Books. Ex. 62(a) (M. Trinh 30(b)(6) Tr. (Jan. 29, 2025)) at 56:13-58:8, 59:23-60:8; C. Blythe Opp. Decl. Ex. 6 (M. Trinh 30(b)(6) Tr. (July 25, 2025)) at 81:5–82:17, 124:15–25. Microsoft defers to OpenAI.

J. Throughout, OpenAI Knew LibGen Was "Sketchy [as F---]"

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.	329. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Defendants do not dispute that the witnesses testified to the testimony cited below when considered in the full context of their testimony. The cited support references "*infra* §§ II.J.1-2," which does not itself constitute evidence; to the extent the facts asserted in those sections are individually disputed, so too is this summary assertion. Further, the Court should disregard Paragraph 329 because the assertion that "piracy is wrong and illegal" constitutes improper attorney argument and/or is conclusory. It thus requires no response.

1. OpenAI's Contemporaneous Knowledge

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	330. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 220 contains the quoted language. However, disputed that Smith "informed"

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.*	Brockman of these facts in the manner Plaintiffs suggest. Smith shared boilerplate language about LibGen and Sci-Hub in the context of attempting to post a link to an article about Brockman on his Wikipedia page; the exchange does not reflect a substantive warning or briefing regarding OpenAI's use of LibGen data. Books Ex. 220 at OPCO_MDL_PD_000003022. Further disputed because Brockman could not recall this conversation during his deposition, nor could he recall reading this message. C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 43:25–49:2. Microsoft defers to OpenAI.
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.	331. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 9 contains the quoted testimony. However, disputed that Plaintiffs' characterization that Brundage "acknowledged" that his legal concerns about OpenAI's activities included "copyright concerns" accurately reflects his testimony; Brundage testified [redacted]. C. Blythe Opp. Decl. Ex. 32 (M. Brundage 30(b)(1) Tr.) at 273:21–274:17. Further disputed to the extent Plaintiffs suggest that copyright concerns were the legal concerns Brundage had in mind; Brundage testified [redacted]. *Id.* Further disputed that Brundage was OpenAI's "Head of Policy Research" in 2019; Brundage [redacted]

	[redacted]. C. Blythe Opp. Decl. Ex. 32 (M. Brundage 30(b)(1) Tr.) at 42:4–11. Microsoft defers to OpenAI.
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.	332. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 785 contains the quoted testimony. However, disputed to the extent Plaintiffs' phrasing suggests that Wu "stated" this in 2019. Wu testified [redacted]. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 91:10–25. Microsoft defers to OpenAI.
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.	333. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 337 contains Chess's quoted statement. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Chess's reference to "legal issues" suggest that he was referring to copyright issues. Books Ex. 337 provides that Ben Chess was to "look at the different possible data sets" and to "see if any meet the criteria we're looking for" and, in response to that ask, stated that he doesn't "have a lot of context on what the criteria is" and acknowledged that he was aware generally of legal issues, "but only caught those details in

	passing." Books Ex. 337 at OPCO_MDL_004606758. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	334. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Exs. 384 and 809 contain the quoted statements. However, disputed that Plaintiffs' characterization that Amodei "pasted" the text from the LibGen Wikipedia entry is supported by the cited evidence. Amodei testified [redacted] C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 155:11-157:19. Further disputed to the extent Plaintiffs' characterization suggests that Amodei personally reviewed or obtained the quoted information from the LibGen Wikipedia page because the cited evidence does not establish that proposition. *Id.* Microsoft defers to OpenAI.
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.	335. **Undisputed.** Microsoft defers to OpenAI.
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	336. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that on April 18, 2019, Mr. Altman shared the April 2019 document with

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." This document was also shared with other executives at Microsoft, including Chief Technology Officer Kevin Scott. *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. Nelson Decl. Ex. 702; Nelson Decl. Ex. 703 at -984–85.	Mr. Gates; under the heading "Data," the document states that OpenAI "added another ~11B words from Library Genesis (LibGen), which contains books, textbooks, science magazine articles, and similar context," and reports that the model grew in size from GPT-2; and the document was also shared as an email attachment with two Microsoft employees (Chief Technology Officer Kevin Scott and his technical advisor Umesh Madan). Defendants dispute the remaining portions because the cited evidence does not support the assertion. Defendants dispute that the April 2019 document supports that the model "had grown nearly tenfold in size from GPT-2" That document states that GPT-2 was trained on "~6B words." An addition of "~11B words" from LibGen would not constitute a "tenfold increase in size." Books Ex. 703 at MSFT_AI_MDL_000007984–MSFT_AI_MDL_000007985; Books Ex. 812 at OPCO_MDL_002135911. Microsoft also responds as follows: Microsoft disputes any implication in Paragraph 336 that Bill Gates was a Microsoft executive acting on Microsoft's behalf and/or with decision-making authority over Microsoft's decision to invest in OpenAI. *See* Response to Paragraph 187.
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	337. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants do not dispute that Books Exs. 670, 384, and 385 contain the quoted language. But the facts set forth in Paragraph 337 are disputed on the

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*.	grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Microsoft also disputes any implication in Paragraph 337 that any Microsoft witness who received the April 2019 document (Books Ex. 812 and Books Ex. 703) followed the hyperlink to the Wikipedia page describing LibGen. *See* Response to Paragraph 190. *See* C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:17-22; C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 118:19-121:16.
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.	338. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 4 contains the quoted text. However, disputed to the extent Plaintiffs' characterization implies that Amodei testified that anyone at OpenAI in fact read the Wikipedia article. In response to Plaintiffs' counsel's question that [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 159:7-14. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	339. **Disputed in Part.** OpenAI responds as follows: Undisputed that the link was sent in the Slack channel. However, disputed to the extent Plaintiffs' characterization implies that Schulman was recommending using LibGen or referencing the specific contents of LibGen. Schulman was sharing a UI design that OpenAI could use as inspiration for an internal website where OpenAI would host tech reports written by employees. Nothing in this Slack message [redacted] C. Blythe Opp. Decl. Ex. 30 (J. Schulman 30(b)(1) Tr.) at 103:19-106:14. Microsoft defers to OpenAI.
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.	340. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited exhibit contains the quoted statement. However, disputed that Schulman "recognized the legal exposure that LibGen created." Plaintiffs' characterization is inaccurate and incomplete because the cited document shows that Schulman was responding to a message from Che Chang, an attorney for

	OpenAI, who was providing privileged legal advice. Schulman also testified that he guessed [redacted] C. Blythe Opp. Decl. Ex. 30 (J. Schulman 30(b)(1) Tr.) at 149:12–150:3. Microsoft defers to OpenAI.
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.	341. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited exhibit contains the quoted statements. However, disputed that Plaintiffs' characterization of and inferences drawn from those statements are corroborated by the cited evidence. When Amodei [redacted] C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 178:10–181:6. Microsoft defers to OpenAI.
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.*	342. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited exhibit contains the quoted statements. However, disputed that Chess wrote the comment or "questioned the legality of LibGen." The cited evidence does not support Plaintiffs' assertion that Chess authored the comment. Chess testified that [redacted]

	[redacted]. C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) at 134:16–138:20. Further, when Plaintiffs questioned Chess [redacted] C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) at 59:7–13. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.	343. **Undisputed.** Microsoft defers to OpenAI.
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.	344. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed that the cited evidence supports Plaintiffs' assertion. Plaintiffs added an exclamation mark to the end of the quoted statement where one did not exist in the underlying document. *See* Books Ex. 302 at OPCO_MDL_001937705. Further, disputed that Plaintiffs' characterization of the statement as a backward-looking assertion that OpenAI had already "trained GPT-3 on pirated stuff" is supported by the record. GPT-3 had not been trained as of August 19, 2019,

	and Mann testified that the statement was "[redacted]" for that reason. C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 83:14-84:20. Mann further testified that [redacted]. *Id.* at 86:4-87:14. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.	345. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Exs. 816 and 841 contain the quoted language. However, disputed to the extent Plaintiffs suggest that OpenAI was aware that "the project is in blatant violation of copyright." Chess testified that [redacted]" C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) at 118:18–120:4. Microsoft defers to OpenAI.

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.	346. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the metadata listed Tom Brown as the custodian of a screenshot bearing the referenced statement. However, disputed that Mr. Brown was aware of the entirety of the words in this screenshot, that he took the screenshot, and that he intentionally placed the screenshot in his personal files. Mr. Brown testified that [redacted] *See* C. Blythe Opp. Decl. Ex. 34 (T. Brown 30(b)(1) Tr.) at 35:25-40:5. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	347. **Disputed in Part.** OpenAI responds as follows: Undisputed that Mann gave the quoted testimony. However, disputed that Plaintiffs' characterization of Mann's testimony is accurate or complete. Plaintiffs' juxtaposition of different portions of Mann's deposition testimony misrepresents the timeline and Mann's knowledge. First, Mann testified [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 260:17-261:1. Plaintiffs excerpt Mann's later testimony—in which he stated [redacted]—without including the question that elicited it, in which Plaintiffs' counsel asked [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 272:20-273:2. Second, [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 277:5-8. Microsoft defers to OpenAI.
348. Mann testified that "LibGen was sketchy [as f---]." *Id.* at 77:18–78:15.	348. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 33 contains the quoted language "LibGen is sketchy AF." However, disputed that Plaintiffs accurately characterize Mann's testimony. Plaintiffs took the quote out of context. During his deposition, Plaintiffs' attorney, Joseph Saveri, asked Mann to [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 77:10-78:18. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.

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.	349. **Disputed in Part – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 380 contains the quoted language. However, disputed that Plaintiffs' characterization of, and inferences drawn from, that language are accurate or complete, because Plaintiffs omit context necessary to understand it. Specifically, Plaintiffs present Hallacy's draft as though it reflects a corporate position on data sourcing, but a statement by a single employee is insufficient to establish the underlying fact, at all times and for all purposes in this litigation, including as to any belief or understanding held by OpenAI. Books Ex. 380 is a personal draft document by Hallacy, not an approved corporate policy, directive, or communication to leadership. See Books Ex. 380 at OPCO_SDNY_1636049, OPCO_SDNY_1636055. Additionally, Hallacy's tracked changes from "legit sources" to "commercially licensed sources" reflect a development in thought, not any acknowledgment. Books Ex. 380 at OPCO_SDNY_1636049. The statement that "[i]t's also not clear if we can commercialize products obtained via scraping" was an open question about scraping generally, not a specific statement about LibGen as suggested by the sequencing and implied connection in Plaintiffs' presentation. Books Ex. 380 at OPCO_SDNY_1636049, OPCO_SDNY_1636055. Finally, the document's forward-looking framing—"Once AGI is hit we're gonna want to commercialize"—shows Hallacy was planning for future

	data needs, not commenting on the status of existing datasets. Books Ex. 380 at OPCO_SDNY_1636049. Microsoft defers to OpenAI. Microsoft further responds that, to the extent Paragraph 349 is offered to establish what Microsoft wanted or intended, the facts are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is hearsay and lacks foundation; it is a draft by an OpenAI employee about what "MSFT" wants.
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.	350. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 230 contains the quoted statement attributed to Hallacy. However, disputed that Book Plaintiffs' characterization of, and inferences drawn from, the quoted statement are supported by the cited evidence. Specifically, Book Plaintiffs characterize the processing step as designed to "eliminate copyright info," but Hallacy's comment describes a routine data-cleaning step to remove boilerplate text from the beginning of documents in order to improve data quality for model training. *See* Books Ex. 230 at OPCO_MDL_000135864. Further, Book Plaintiffs overstate the testimony by claiming "Hallacy participated in developing that processing step," but the cited testimony shows that Hallacy testified about reviewing and commenting on the pull request, not

	authoring or developing it. *See* Books Ex. 25 (C. Hallacy 30(b)(1) Tr.) at 117:19–120:14. Microsoft defers to OpenAI.
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.	351. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 818 contains the quoted language. However, disputed that Plaintiffs' characterization of, and inferences drawn from, the quoted language are supported by the cited evidence. Specifically, Amodei's statement is aspirational, expressly noting that he did not know if such a technical capability existed. The message reflects Amodei raising a research question about machine unlearning, not an admission that OpenAI's existing use of training data was unlawful. *See* Books Ex. 818 at OPCO_MDL_001939125-OPCO_MDL_001939126. Microsoft defers to OpenAI.
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.	352. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 819 contains the quoted language. However, disputed that the quoted language carries the meaning Plaintiffs imply. Book Plaintiffs omit context necessary to understand or evaluate the quoted language. Specifically, Mann testified that [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 77:10-78:18. Amodei [redacted] C.

	Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 213:18-216:6. Microsoft defers to OpenAI.
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.*	353. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 819 contains the quoted phrases. However, disputed that the cited evidence supports Plaintiffs' characterization of an unattributed bullet point in meeting notes as an "OpenAI direct[ive]." The document is meeting notes, and it is not established who made the notation, whether it was a directive, a proposal, a discussion point, or a personal note. Characterizing an unattributed bullet point in meeting notes as a corporate "direct[ive]" is unsupported by the cited evidence. Further, Plaintiffs juxtapose "no code, vague on data" with "Libgen is sketchy [as f---]" to imply that the vagueness was motivated by a desire to conceal LibGen's use, but the cited evidence does not establish that connection. Mann testified that [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 409:7-16. The "vague on data" notation is consistent with this safety rationale, not concealment of LibGen. The fact that both phrases appear in the same

	meeting notes does not establish a causal or logical relationship between them. Microsoft defers to OpenAI.
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.	354. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 819 contains the referenced phrase. However, disputed that Plaintiffs' characterization of the cited evidence is accurate or complete. Specifically, Plaintiffs state that "OpenAI researchers, including Tom Brown, described LibGen as 'sketchy AF,'" but Books Ex. 819 is a set of meeting notes that do not attribute the phrase "sketchy AF" to Tom Brown or to any specific individual. Plaintiffs have cited no evidence establishing that Brown authored or spoke this particular notation. Further, Plaintiffs characterize the notation as the view of "OpenAI researchers" — plural — but the meeting notes contain a single unattributed line, and there is no basis to attribute it to multiple researchers. Additionally, in context, "sketchy" [redacted]. *See* C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 77:10-78:18. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
355. In July 2020, another OpenAI employee wrote: "it does seem nice to try to buy the books, and maybe	355. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 246 contains the quoted

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.	language. However, disputed that Plaintiffs' characterization of, and inferences drawn from, the quoted statements suggest that they reflected OpenAI decisions concerning whether to identify the source of books used in training data. Plaintiffs identify the speaker only as "another OpenAI employee" and omit the context of the discussion, including what prompted the statements and whether the statements reflected an individual suggestion or any decision by OpenAI. A statement by a single employee, presented without context, is insufficient to establish OpenAI's point of view or practices concerning the source or disclosure of book data. Microsoft defers to OpenAI.
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.*	356. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 325 contains the quoted statement. However, the facts set forth in Paragraph 356 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is irrelevant and immaterial to resolution of Plaintiffs' Motion under Fed. R. Evid. 401 and 402 because the "book-summarization project" involved a model that is not an in-scope model and is not subject to this litigation. *See* ECF 707 (Stein Opinion & Order re Models and MTS in Class). Further, Plaintiffs' characterization of the purpose of "the book-summarization project" is disputed. The purpose of

	the project was to [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 121:23–122:8; 126:19–127:23; C. Blythe Opp. Decl. Ex. 35 at OPCO_MDL_000000069-70. Microsoft defers to OpenAI.
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.	357. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Nelson Decl. Ex. 325 contains the quoted statement. However, the facts set forth in Paragraph 357 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the cited material is a continuation of the discussion from Paragraph 356 about the "book-summarization" project, which involved a model that is not an in-scope model and is not subject to this litigation. *See* ECF 707 (Stein Opinion & Order re Models and MTS in Class). The evidence cited is therefore irrelevant and immaterial to resolution of Plaintiffs' Motion under Fed. R. Evid. 401 and 402. Further, Plaintiffs' characterization that the conversation concerned "OpenAI's policy towards training on copyrighted data" is inaccurate. The purpose of the discussed project did not concern policies on "training on copyrighted data" but rather [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 121:23–122:8; 126:19–127:23; C. Blythe Opp. Decl. Ex. 35 at OPCO_MDL_000000069-70. Plaintiffs quote the thread out of context, and the inferences Plaintiffs draw therefrom are not corroborated by the cited evidence. *See* Books Ex. 325 at OPCO_MDL_004413315. Microsoft defers to OpenAI.
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.	358. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that Nelson Decl. Ex. 223 contains the quoted statement. However, the facts set forth in Paragraph 358 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, this thread reflects a further discussion about the "book summarization" project noted in Paragraph 356, which involved a model that is not an in-scope model and is not subject to this litigation. *See* ECF 707 (Stein Opinion & Order re Models and MTS in Class). The evidence cited is therefore irrelevant and immaterial to resolution of Plaintiffs' Motion under Fed. R. Evid. 401 and 402. The purpose of the underlying project was to [redacted]

	[redacted] Ex. 23 (J. Wu 30(b)(1) Tr.) at 121:23–122:8. Microsoft defers to OpenAI.
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.	359. **Undisputed.** Microsoft defers to OpenAI.
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.*	360. **Disputed in Part – Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that 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. *See* Books Ex. 276 at OPCO_MDL_001658034. However, OpenAI disputes Plaintiffs' reliance on the reporter's inquiry to the extent it is offered to prove the truth of any matter asserted or assumed in the reporter's questions, including how GPT-3 was trained, whether the public had access to the contents of Books1 and Books2, or how OpenAI assembled them, because the reporter's questions as set forth in the first sentence constitute inadmissible hearsay. Because the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, it is therefore is not properly before the Court. Microsoft defers to OpenAI.

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.*	361. **Undisputed.** Microsoft defers to OpenAI.
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.*	362. **Disputed in Part**. OpenAI responds as follows: Undisputed that the quoted text appears in the cited document. However, disputed as to Plaintiffs' implication that McAleese was referring to OpenAI. McAleese's comment was a response to OpenAI employees speculating about the actions of Mistral, another company. After speculating that Mistral may be training on all of LibGen, another OpenAI employee asked, "Mistral is open source tho [sic], so shouldn't it be quite easy to uncover that?" It was only after this exchange that McAleese stated, about Mistral: "Train on libgen, raise money, get new data, delete the old models, be clean forever more. Temporal regulatory arbitrage." *See* Books Ex. 324 at 5-6. Microsoft defers to OpenAI.
363. Another OpenAI employee in the same channel asked, "What's libgen?" Nelson Decl. Ex. 324 at -	363. **Undisputed.**

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.*	Microsoft defers to OpenAI.
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.*	364. **Disputed in Part.** OpenAI responds as follows: Undisputed that 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." However, disputed to the extent Paragraph 364 implies that the removal of references was intended to conceal the use of Library Genesis materials. Wu testified [redacted]. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 95:23-96:9. Microsoft defers to OpenAI.
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	365. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Chung received the cited OpenAI paper or that he gave the quoted testimony. But Microsoft disputes that Chung "forwarded it to others within Microsoft." Chung testified that, based on the email shown to him, he "sent it to one person" and had no recollection of who else he sent it to. C. Blythe Opp. Decl. Ex. 36 (E. Chung 30(b)(1) Tr.) at 61:10–25. Microsoft

responded that he didn't "have any feelings about it." Nelson Decl. Ex. 14 at 60:20–61:15.	further disputes any implication that Chung understood the paper to show that OpenAI trained on LibGen. He testified that he did not "think it's possible to infer any conclusion" from the draft's reference to "train/test" datasets. *Id.* at 59:25-60:16. Microsoft further responds that the facts set forth in Paragraph 365 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. Further, Chung's lack of "feelings" about OpenAI's later publication choices has no bearing on any such claim or defense.
366. Ouyang described the LibGen data as "pirated books." Nelson Decl. Ex. 348.	366. **Undisputed.** Microsoft defers to OpenAI.
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.	367. **Undisputed.** Microsoft defers to OpenAI.

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.	368. **Disputed in Part**. OpenAI responds as follows: Undisputed that Books Ex. 815 contains the quoted text. However, disputed that Plaintiffs' characterization accurately reflects Xu's testimony concerning LibGen. When asked [redacted] C. Blythe Opp. Decl. Ex. 56 (T. Xu 30(b)(1) Tr.) at 23:11-24:13. Xu only stated that LibGen [redacted] C. Blythe Opp. Decl. Ex. 56 (T. Xu 30(b)(1) Tr.) at 24:15-22. Further disputed because Plaintiffs' characterization of Xu's statement that "in general, you shouldn't" is a misleading characterization of his testimony. That statement was elicited in response to a hypothetical concerning [redacted] C. Blythe Opp. Decl. Ex. 56 (T. Xu 30(b)(1) Tr.) at 83:16-84:5. This hypothetical did not concern LibGen specifically and was not tied to OpenAI's conduct. Microsoft defers to OpenAI.

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.	369. **Disputed in Part.** OpenAI responds as follows: Undisputed that Nelson Decl. Ex. 55(b) contains the quoted testimony. However, OpenAI further responds that the facts set forth in Paragraph 369 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning fairness generally are not facts and are immaterial to the issues before the Court. Microsoft defers to OpenAI.
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.	370. **Undisputed.** Microsoft defers to OpenAI.
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.* 168:17–23.	371. **Undisputed.** Microsoft defers to OpenAI.
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	372. **Disputed.** OpenAI responds as follows: Paragraph 372 mischaracterizes Altman's testimony. Altman did not make any affirmative statement about what

initial act of piracy would be "illegal" on its own. *Id.* at 167:8–14.	conduct would or would not be "appropriate." Rather, Altman was responding to Plaintiffs' counsel's question: "Do you think it would be appropriate to instead of buying a copy of that book, go to a pirated website, download that copy of the book, and then write a review of it?" Altman responded, "No. I would choose to go buy a copy of that or borrow a copy of that book." Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 167:6-14. Further, Altman did not characterize obtaining a book as an "initial act of piracy." Plaintiffs' counsel asked Altman, "[Y]ou understand that piracy is illegal?" Altman responded, "I do." Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 167:15-17. Paragraph 372 thus improperly attributes to Altman characterizations and conclusions that originated with Plaintiffs' counsel, not with the witness. Microsoft defers to OpenAI.
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.	373. **Undisputed.** Microsoft defers to OpenAI.
374. Daniela Amodei, OpenAI's former Vice President of Safety and Policy, testified that it is not	374. **Undisputed.**

appropriate "to go to an illegal pirate website and take data from there for free." Nelson Decl. Ex. 3 at 43:11–15.	Microsoft defers to OpenAI.
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.	375. **Disputed in Part.** OpenAI responds as follows: Undisputed that Chess testified he "would believe" piracy is "illegal." However, disputed that Chess testified that "downloading copyrighted material . . . for free" is illegal. Chess [redacted] C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) at 20:21–25. Further disputed that Chess "had downloaded copyrighted content from The Pirate Bay." Chess did not confirm that he downloaded any content from The Pirate Bay, testifying instead that [redacted] *Id.* at 91:13–22. Microsoft defers to OpenAI.
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.	376. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 58 contains the quoted testimony. However, disputed to the extent Paragraph 376 presents Sutskever's testimony in a misleading manner by omitting context necessary to understand the quoted language. Specifically, Sutskever's reference to "a dataset like the one that we're discussing" [redacted] C. Blythe Opp. Decl. Ex. 55 (I. Sutskever 30(b)(1) Tr.) at 151:7-13. Microsoft defers to OpenAI.

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.	377. **Undisputed.** Microsoft defers to OpenAI.
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*.	378. **Disputed in Part – Not Supported by Admissible Evidence; Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 821 contains the quoted statements. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those statements suggest that they reflect OpenAI's views. The statements were made by Gogineni in a personal social media exchange with third parties and are not statements by OpenAI. The disputed proposition is therefore not supported by the cited evidence. Further disputed to the extent Plaintiffs rely on statements by third parties in the exchange for their truth; those statements constitute inadmissible hearsay. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Because the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, it is therefore not properly before the Court. Microsoft defers to OpenAI.
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.	379. **Disputed in Part – Not Supported by Admissible Evidence; Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 822 contains the quoted statement. However, disputed that Plaintiffs' characterization of the Teknium

	account as belonging to Gogineni's "colleague" is supported by the cited evidence; neither the exhibit nor the cited deposition testimony establishes that the individual was Gogineni's colleague. The disputed proposition is therefore not supported by the cited evidence. Further disputed to the extent Plaintiffs' characterization of, and inferences drawn from, this personal social media exchange suggest that it reflects OpenAI's views. Also disputed to the extent Plaintiffs rely on statements by third parties in the exchange for their truth; those statements constitute inadmissible hearsay. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Because the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, it is therefore not properly before the Court. Microsoft defers to OpenAI.
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.	380. **Disputed in Part.** OpenAI responds as follows: Undisputed that Gogineni stated that it was "impossible" to find a copy of a book either "legally or illegally." However, disputed that the statement concerned a book Gogineni personally "wanted," as that characterization is not reflected in the cited tweet or related deposition testimony. Books Ex. 823. Microsoft defers to OpenAI.

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.	381. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 33 contains the quoted testimony. Undisputed that 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. However, disputed that Mann was "the OpenAI employee who torrented the LibGen collections," as the cited deposition testimony does not establish that fact. Disputed that Mann testified that he downloaded "books, movies, and software" from The Pirate Bay. Mann acknowledged downloading movies, but testified [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 232:16-233:6. Disputed that Mann "agreed" that "downloading copyrighted material without permission [could] be a crime"; when presented with that proposition, Mann responded, [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr. (Mar. 3, 2025)) at 233:16-20. For those reasons, the disputed propositions are not supported by the cited evidence. Microsoft defers to OpenAI.
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,	382. **Undisputed – Immaterial; Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Mr. Nadella made or agreed with those statements at his deposition. But, pursuant to Fed. R. Evid. 106, Mr.

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.	Nadella’s full testimony was: “Q. And, sir, this is, again, another article. This is now from 2018, April of 2018. And just that first sentence, this is a Microsoft blog, states that: Software piracy and fraudulent subscriptions are serious, industry-wide problems affecting consumers and organizations around the world. You agree with that, sir, correct? A. I do. Q. And by the way, we’re talking about it in terms of software piracy, but the same piratical problem exist, for example, when there is something with respect to movies, correct? A. Yeah. All intellectual property, yeah. Q. Yes. And the same issues with respect to piracy occur with respect to TV shows, fair? A. Yeah. Q. And the same issues with respect to piracy occur with respect to music? A. Yes. Q. And the same issues with respect to piracy occur with respect to books? A. Sure. Q. In other words, it’s not like software stands apart from the other categories of intellectual property that exist, correct? A. They do. It’s the same intellectual—on the piracy side it’s the same, we have a cyber issue, which is separate. Q. And you know that anybody who visits one of these sites and downloads pirated copyright material is at risk of legal liability, correct? A. Yeah, I think, as I understand it, if I look at all these pirated or mirror sites, it is – people do this even to save CDN costs. Because it’s public domain stuff, may have done -- done through terms of service even with scraping. And so there’s a lot of that that happens in the -- in the industry that is not -- even if it’s public domain that’s available, people have these mechanisms to save money on download, for example, and replication and so on. And so those are all activities that

	are sort of on the spectrum of what is legal and illegal, which I think is to be litigated by lawyers." Books Ex. 40 (S. Nadella 30(b)(1) Tr.) at 172:21-175:12. Regarding the quote that Microsoft should "not use pirated sites to obtain data," Mr. Nadella also testified: "So, again, the two sides of things. One is copyrights have to be protected. Second is if the data is available publicly, using that data to create a transformative technology that is not a regurgitation engine, that's fair use. That's sort of how all of software, every era of software has gotten created…You've shown me throughout today that LibGen is a pirated site and it had content. And I have no idea what, exactly, it had. Did it have public domain data? Did it have only copyrighted data? What data from there was actually used in training runs? I have none of that information. So therefore, I would stick with – the principle I would still stand by is that which is, you know, that people should just use public data for training, not use pirated sites to obtain data, and that they should adhere to robots.txt norms." Books Ex. 40 (S. Nadella 30(b)(1) Tr.) at 298:16-21, 300:1-11. Microsoft further responds that the facts set forth in Paragraph 382 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, copyright "piracy" in general is not the issue before the Court. OpenAI defers to Microsoft's response.

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.	383. **Undisputed – Immaterial; Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Mr. Nadella agreed with the quoted statements in Paragraph 383 at his deposition. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates Mr. Nadella's full testimony stated in response to Paragraph 382. Microsoft further responds that the facts set forth in Paragraph 383 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, copyright "piracy" in general is not the issue before the Court. OpenAI defers to Microsoft's response.
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.*	384. **Undisputed – Immaterial; Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Mr. Nadella made the statements quoted in Paragraph 384 at his deposition. But, pursuant to Fed. R. Evid. 106, Mr. Nadella further testified: "Q. I'm going to assume-have you assume for a second here that, in fact, LibGen was used in the training of certain OpenAI models. You with me?...A. You are basically saying I should make an assumption that it was used. Q. Yes. A. Okay. Q. Assuming that it was used, assuming that LibGen was used, do you regret that content from pirated data with copyrighted material was used to train OpenAI's large language models on Microsoft systems? A. I would say two things that come to me when you say that. One is, I wish we had not -- or OpenAI had not

	gone to a pirated site. They could have scanned the copy, they should have done whatever --because I sort of say whatever it is that you needed, you could have used it as fair use. And so therefore, do what Google did with Google Books. Go on library read it. Scan it. So that's what I would have done. So I -- so to your question, I would separate the two. I think I would regret using a pirated site, but I would not regret using a book as fair use to create a transformative technology." Books Ex. 40 (S. Nadella 30(b)(1) Tr.) at 302:8-303:1. Microsoft further responds that the facts set forth in Paragraph 384 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, copyright "piracy" in general is not the issue before the Court. OpenAI defers to Microsoft's response.
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.	385. **Undisputed – Incomplete; Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Gates made the quoted statements in Paragraph 385 at his deposition. But, pursuant to Fed. R. Evid. 106, Mr. Gates qualified that testimony: "If you're *copying something exactly* and using it without paying for it, I think that's right." C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 19:24-20:1 (emphasis added). Elsewhere, Mr. Gates testified, "If you're making fair use of material, I'm not sure

	the source matters. But, you know, that's kind of a legal question." *Id.* at 118:4-6. Microsoft further responds that the facts set forth in Paragraph 385 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy" generally are not facts and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	386. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Madan agreed to the first statement in Paragraph 386, or that he made the quoted statements in the second sentence at his deposition. Microsoft also responds that the facts set forth in Paragraph 386 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy" generally are not facts and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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	387. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr.

material from pirated websites. Nelson Decl. Ex. 61 at 35:16–36:4.	Tiwary agreed to the statement in Paragraph 387 at his deposition. Microsoft also responds that the facts set forth in Paragraph 387 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy" generally are not facts and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	388. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Suleyman made the statements quoted in Paragraph 388 at his deposition. Microsoft also responds that the facts set forth in Paragraph 388 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, the training of Microsoft's own LLMs is immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	389. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Usdan made the statements quoted in Paragraph 389 at his deposition.

	Microsoft further responds that the facts set forth in Paragraph 389 are immaterial because these facts do not affect the outcome of the suit under the governing law. The training of Microsoft's own LLMs is immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	390. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Bilenko made the statements quoted in Paragraph 390 at his deposition. Microsoft further responds that the facts set forth in Paragraph 390 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy" generally are not facts and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	391. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Hecht made those statements in Paragraph 391 at this deposition. Microsoft further responds that the facts set forth in Paragraph 391 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy"

	generally are not facts and are immaterial to the issues before the Court. Further, a Microsoft employee's personal views on whether he would download content from a particular website have no bearing on any element of the claims or defenses in this case. Hecht also testified that he has no direct knowledge of OpenAI's use of LibGen, C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 374:19–24, and that he is not "the arbiter of what's pirated or isn't," *id.* at 373:18-19. OpenAI defers to Microsoft's response.
392. [redacted]	392. **Undisputed – Immaterial; Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Books Ex. 151 contains the quoted text. But, pursuant to Fed. R. Evid. 106, [redacted] Books Ex. 151 at MSFT_AI_MDL_000004681. Microsoft further responds that the facts set forth in Paragraph 392 are immaterial because these facts do not affect the outcome of the suit under the governing law. Microsoft's general employee code of conduct is immaterial to the issues before this Court. OpenAI defers to Microsoft's response.

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.	393. Undisputed that The Pile is a training dataset developed by Leo Gao, Shawn Presser, and others, and that The Pile includes a component called "Books3," which is a dataset of books derived from a copy of the contents of the Bibliotik dataset. Microsoft defers to OpenAI.
394. Presser stated that Books3 derived from "all of Bibliotik" and contained 196,640 books. Nelson Decl. Ex. 307.	394. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 394 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is inadmissible hearsay because this assertion entirely relies on an out-of-court statement by Presser, a non-party whom Book Plaintiffs did not depose. Microsoft defers to OpenAI.
395. The copyright holders of those books did not consent to their inclusion on Bibliotik. Nelson Decl. Ex. 20 at 24:22–25:20.	395. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: This fact is disputed because it is not supported by the cited evidence. Book Plaintiffs rely on deposition testimony from one individual about his subjective belief on whether copyright

	holders of books in the Books3 dataset consented to their inclusion. C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 25:16–20. Microsoft defers to OpenAI.
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.	396. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Gao was a member of EleutherAI and was the first-listed author on the paper announcing The Pile. However, disputed to the extent Paragraph 396 suggests that Gao was the only person who compiled The Pile. The Pile was a joint project among several individuals. Books Ex. 672 at 21 (describing the contributions of the other 11 authors). Disputed that Gao referred to the books in Books3 as "yarrharr data." C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 102:21–103:20. The cited evidence does not support that characterization. Microsoft defers to OpenAI.
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.	397. **Disputed in Part.** OpenAI responds as follows: Undisputed that Mr. Gao so testified. Books Ex. 730; C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 83:18–23, 133:15–22. However, disputed to the extent Book Plaintiffs suggest that using books from Books1 required a license, since any purported use of such books was a fair use. Microsoft defers to OpenAI.

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.	398. **Undisputed.** Microsoft defers to OpenAI.
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).	399. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 399 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is inadmissible hearsay, as this assertion relies entirely on a post from X.com by a non-party whom Book Plaintiffs never deposed. Microsoft defers to OpenAI.
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*.	400. **Undisputed.** Microsoft defers to OpenAI.
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	401. **Undisputed.** Microsoft defers to OpenAI.

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.	402. **Undisputed.** Microsoft defers to OpenAI.
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.	403. **Disputed in Part.** OpenAI responds as follows: Undisputed that 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. However, disputed to the extent Paragraph 403 suggests that OpenAI had any role in developing The Pile or downloaded The Pile. Leo Gao was not an employee at OpenAI when EleutherAI developed The Pile and OpenAI did not download a full copy of Books3 or The Pile. C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 185:16-21; S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 78–111. Microsoft defers to OpenAI.
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.	404. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 404 are disputed because they are not supported by the cited evidence. OpenAI did not download

	a full copy of Books3 or The Pile. S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 78–111. Microsoft defers to OpenAI.
405. Microsoft employees were also aware of The Pile. On October 26, 2020, Microsoft employee William Buchwalter posted in a Microsoft Teams message that [redacted] Nelson Decl. Ex. 201.	405. **Disputed in Part – Incomplete.** Microsoft does not dispute that Microsoft employee William Buchwalter exchanged a Teams message with another employee in which "The Pile" was mentioned. But Microsoft disputes Paragraph 405 to the extent it implies that Mr. Buchwalter's "aware[ness] of The Pile" constitutes awareness of the sources of its contents. Pursuant to Fed. R. Evid 106, Microsoft further responds that, when testifying about the document cited as Books Ex. 201, Mr. Buchwalter stated that he had never seen OpenAI's Books2 dataset, he did not know whether OpenAI ever provided Microsoft access to the Books2 dataset, he did not know the source of the Books2 dataset, and he did not know whether the Books2 dataset included books acquired from the alleged sources. C. Blythe Opp. Decl. Ex. 38 (W. Buchwalter 30(b)(1) Tr.) at 45:7–46:25. Microsoft also disputes any implication that Mr. Buchwalter was comparing the sources from which the materials used for The Pile dataset and OpenAI's Books-2 dataset were acquired, or the contents of such materials. Pursuant to Fed. R. Evid. 106, Mr. Buchwalter testified that The Pile dataset "was thought to have something of similar size to what OpenAI was." C. Blythe Opp. Decl. Ex. 38 (W. Buchwalter 30(b)(1) Tr.) at 49:1-2; *see also* Books Ex. 201 (comparing approximate

	number of tokens in The Pile with public information regarding number of tokens used for GPT-3). OpenAI defers to Microsoft's response.
406. Microsoft employee Brandon Norick responded to Buchwalter: [redacted] Nelson Decl. Ex. 201. Buchwalter confirmed the exchange at his deposition. Nelson Decl. Ex. 10 at 53:19–56:5.	406. **Disputed in Part – Incomplete.** Microsoft does not dispute that Books Ex. 201 contains the quoted language. Pursuant to Fed. R. Evid. 106, Microsoft further responds that Mr. Norick testified about his statements in the document cited as Books Ex. 201, and stated as follows: "Q: And what did you mean, considering the source? A. I guess the pile. Q. And why, considering that the source was the pile, did you need more discussion regarding this data? A. I'm not sure." C. Blythe Opp. Decl. Ex. 39 (B. Norick 30(b)(1) Tr.) at 85:22–86:3. Mr. Norick also testified that he did not know the source of the Books3 dataset from the Pile. *See id.* at 84:21–86:3. OpenAI defers to Microsoft's response.
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.	407. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Mr. Gao testified that Jeff Wu participated in the Discord channel for EleutherAI. *See* C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 161:25–165:24. However, disputed that "OpenAI employees also learned of The Pile and its contents through the Discord channel that members of EleutherAI used." The disputed proposition is not supported by the cited evidence. Specifically, neither Books Ex. 237 nor Books Ex. 20 at 161:25–165:24 supports

	the assertion that OpenAI employees learned of The Pile and its contents through the EleutherAI Discord channel. Microsoft defers to OpenAI.
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.	408. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Mr. Gao did not work at OpenAI when the paper announcing The Pile was published. *See* Books Ex. 668; C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 12:10–13:5. Undisputed that Mr. Gao joined OpenAI around October 2021. *See* C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 12:10–13:5. Undisputed that Mr. Gao brainstormed potential datasets to use in training, including Books3; however, Plaintiffs' characterization is inaccurate insofar as it suggests that Books3 was ultimately included, because Mr. Gao testified that he "[u]ltimately" did not include "Books3." C. Blythe Opp. Decl. Ex. 37 (L. Gao 30(b)(1) Tr.) at 146:11-12. Disputed that Mr. Gao "understood that using LibGen 'might get him sued.'" The disputed proposition is not supported by the cited evidence. Specifically, Books Ex. 673 at 86 does not support the assertion that Mr. Gao understood that using LibGen "might get [him] sued." Microsoft defers to OpenAI.
409. Gao stated that "book data was really important for language understanding tasks," and OpenAI employees discussed how "Books3 compare[d] to our libgen	409. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that an internal OpenAI paper identified, as a

datasets." Nelson Decl. Ex. 341; Nelson Decl. Ex. 285 at -159.	point of "future investigation," to see how "Books3 compare[s] to our libgen datasets." *See* Books Ex. 285 at OPCO_MDL_001687159. However, disputed that Mr. Gao "stated that 'book data was really important for language understanding tasks'" because Plaintiffs' characterization is misleading. The cited evidence reflects that Mr. Gao stated he "*thought* book data was really important for language understanding tasks." Books Ex. 341 at OPCO_MDL_004618332 (emphasis added). Plaintiffs' omission of the word "thought" improperly transforms a subjective belief into a declarative assertion of fact. Microsoft defers to OpenAI.
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.	410. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Azure storage containers may be used by OpenAI employees to store personal copies of datasets they were working with in order to prepare them for individual research efforts. However, disputed to the extent Plaintiffs imply that all OpenAI employees engage in this practice or that this is the only use of Azure storage containers. Mike Trinh's testimony states that this is one of many possible uses of Azure storage containers. C. Blythe Opp. Decl. Ex. 40 (M. Trinh 30(b)(6) Vol. 3 Tr. (Mar. 3, 2026)) at 16:6–17:21, 18:18–19:19. For that reason, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.

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.	411. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Mr. Misra ran model-training experiments on a small fraction of The Pile for research purposes using models that are not in-scope in this litigation. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 276:21–277:1, 315:3-19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7. Undisputed that at his deposition, Mr. 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." C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 24:16–25:9. However, disputed that OpenAI downloaded The Pile. The cited evidence indicates that on February 1, 2021, Mr. Misra commented on a GitHub pull request that registered in OpenAI's code repository the storage location of a fraction of The Pile in OpenAI. *See* Books Ex. 277 ("Estimated word count: ~ 13_924_091_820"); C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 276:21–277:1; 315:3–19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7; S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 78–111. Further, the disputed proposition is not supported by the cited evidence. Mr. Misra testified repeatedly that [redacted]. C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 35:18–20, 37:4–7, 47:2–8; C. Blythe Opp. Decl. Ex. 41 (A, Paino 30(b)(6) Tr.) at 276:21–277:1; 315:3–19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7.

	Microsoft defers to OpenAI.
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.	412. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that an OpenAI employee [redacted] #6792. However, disputed that Misra downloaded the data into OpenAI's Azure blob storage container because the cited evidence does not support this assertion. Mr. Misra repeatedly testified that [redacted]. C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 35:18–20, 37:4–7, 47:2–8. Mr. Misra confirmed that [redacted]. C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 72:22–73:1. Similarly, Mr. Piano testified that the pull request at issue was not moving data but rather adding configuration to registry files. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 281:2–5. The record evidence establishes that only a small fraction of The Pile was saved to OpenAI's Azure blob storage systems for use in model-training experiments on out-of-scope research models. *Id.* at 276:21–277:1, 315:3–19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7. Microsoft defers to OpenAI.
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	413. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the February 2021 notes of OpenAI's "Reasoning" team contain the quoted language. However,

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.	disputed that “Misra” “stated” the quoted language because Mr. Misra testified [redacted]. C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 46:7–14. Further, the disputed proposition is not supported by the cited evidence. To the extent Plaintiffs’ assertion implies that 825 GB of The Pile was downloaded to OpenAI’s systems, the cited evidence does not support that inference. Mr. Misra testified that the note [redacted] Books Ex. 277 (“Estimated word count: ~ [redacted]); C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 276:21–277:1, 315:3–19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7; S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 78–111. Microsoft defers to OpenAI.
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.	414. **Undisputed.** Microsoft defers to OpenAI.
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.	415. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Mr. Brockman asked “btw what’s our status on testing out The Pile?” and that Mr. Misra replied that he

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.*	"[s]hould have a comparison this week-next" and that the "[i]nitial signs are positive." However, disputed to the extent Plaintiffs contend that Mr. Misra "used" a full version of The Pile for any purpose, or to the extent that Plaintiffs contend that Mr. Misra used a fractional portion of The Pile for anything other than model-training experiments involving out-of-scope research models. The disputed proposition is not supported by the cited evidence. Books Ex. 331 at OPCO_MDL_004584697 describes training an "x1-x6small" model sweep on The Pile, and OPCO_MDL_004584699 refers to an XL model. Mr. Misra's experiments were very small scale, used very tiny models, and would have required very little data to train; he was trying to get a basic understanding of the quality of The Pile datasets towards evals commonly tracked in pretraining and to compare them to baselines on datasets that OpenAI had internally, and this was the full extent of that work. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 215:14–216:2. Microsoft defers to OpenAI.
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	416. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that EleutherAI included [redacted] and [redacted] in The Pile. However, disputed that OpenAI's use of two datasets that EleutherAI also included in The Pile means that OpenAI sourced them from EleutherAI or sourced them from a full download of The Pile. The

[redacted] for training, *see* Nelson Decl. Ex. 312.	disputed proposition is not supported by the cited evidence. The tools that EleutherAI released for accessing The Pile allowed anyone to download individual dataset components without downloading the full Pile. *See* C. Blythe Opp. Decl. Ex. 44 https://github.com/EleutherAI/the-pile. Further, some of those components were available publicly from other sources. In fact, The Pile itself sourced some components, such as [redacted], from others. *See* S. Krishnan Opp. Decl. (Books) Ex. A ¶ 105. Microsoft defers to OpenAI.
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.	417. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Nelson Decl. Ex. 313 indicates that Jeff Wu initiated pull request #47861 titled "download the pile." However, disputed that Jeff Wu downloaded The Pile—or any portion thereof—to his personal container, as Plaintiffs cite no evidence establishing that Mr. Wu downloaded *anything* to his personal container or that Mr. Wu downloaded the full Pile. Further, the facts are disputed because the cited evidence is incorrect. At the time of the alleged download in September 2023—as Plaintiffs admit—"EleutherAI had removed Books3 from its own hosted version of The Pile". Further disputed that EleutherAI removed Books3 from its hosted version of The Pile in August 2023. The relevant change appears to have been made in November 2020. *See* C. Blythe Opp. Decl. Ex. 45, https://github.com/EleutherAI/the-

	pile/commit/71e3a1c026ac2b299a7b07dcf7d56a4d850f7076; *see also* C. Blythe Opp. Decl. Ex. 46, https://github.com/EleutherAI/the-pile/commit/8ef385fb0c5fe5f9fd4bef594ff20fe232a709aa. Microsoft defers to OpenAI.

B. OpenAI Deleted Its Copy of The Pile in October 2024

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.	418. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Misra left OpenAI in 2021. However, disputed that Misra downloaded The Pile into his personal storage account. The cited evidence does not establish that Vedant Misra downloaded The Pile into his personal container. Misra testified repeatedly [redacted] C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) at 35:18–20; 37:4–7; 38:1-13; 47:2-8. Further, the record evidence establishes that only a small fraction of The Pile was saved to OpenAI's Azure blob storage systems for use in model-training experiments on out-of-scope research models. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 276:21–277:1; 315:3–19; C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:7. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.

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.	419. **Disputed – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that the First Consolidated Amended Complaint contains speculation as to the source of OpenAI's "Books2" dataset used to train GPT-3, including speculation that the source was Z-Library or Books3, but it does not make the identification Plaintiffs now attribute to it. First Consolidated Am. Compl. ¶ 44. However, OpenAI disputes Paragraph 419 because it mischaracterizes the First Consolidated Amended Complaint, as it does not identify "The Pile" as stated in Paragraph 419. Further, the First Consolidated Amended Complaint was not filed in "March 2024" as Paragraph 419 asserts. The facts set forth in Paragraph 419 are therefore unsupported by the cited material.
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.	420. **Undisputed.** Microsoft defers to OpenAI.
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.	421. **Disputed in Part.** OpenAI responds as follows: Undisputed that data from Misra's research container was not searched or produced before the container was removed through routine IT maintenance. However, disputed to the extent Paragraph 421 suggests that OpenAI had an obligation to search for data from Mr. Misra's research container, especially given the lack of evidence that any relevant data was saved there. Plaintiffs

	never requested that Mr. Misra be included as an ESI custodian. Plaintiffs did not request any search of any Azure containers until October 16, 2025, and Plaintiffs did not request a search of Mr. Misra's container until November 1, 2025. C. Blythe Opp. Decl. Ex. 47 (J. Connors Oct. 16, 2025 Email); C. Blythe Opp. Decl. Ex. 48 (A. Magee Nov. 1, 2025 Email). This was over one year after routine IT maintenance resulted in removal of unused storage containers assigned to former employees that had not been used in an extended period of time. Microsoft defers to OpenAI.
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.	422. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that, at the time of the removal of accounts belonging to certain former employees, [redacted] *See* C. Blythe Opp. Decl. Ex. 40 (M. Trinh 30(b)(6) Vol. 3 Tr. (Mar. 3, 2026)) at 38:9–25. However, disputed as to Plaintiffs' characterization of Mr. Trinh's testimony. The disputed proposition is not supported by the cited evidence. Mr. Trinh testified that [redacted] *Id.* at 37:2–38:7; *see also id.* at 39:1–7. The cited testimony does not support Plaintiffs' assertion that [redacted]; rather, Mr. Trinh

	testified [redacted]." *Id.* at 37:2–38:7. Microsoft defers to OpenAI.
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.	423. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: disputed as the proposition is not supported by the cited evidence. Specifically, Mr. Trinh did not testify, nor does the cited testimony demonstrate that he was even asked, whether data stored in Mr. Misra's research container in October 2024 was stored anywhere else. C. Blythe Opp. Decl. Ex. 40 (M. Trinh 30(b)(6) Vol. 3 Tr. (Mar. 3, 2026)) at 43:2–8, 46:4–6. Further, Mr. Trinh admitted that he had no personal knowledge of any data deletion or restoration, and that he only heard about it because it was "one example that Mr. Duncan described to me" and that "my understanding is that -- that Mr. Findlay passed along to me[.]" *Id.* at 40:14–16; 41:1–2. When asked whether he was aware of OpenAI making any efforts to restore the deleted data, Mr. Trinh responded: "I would have to check with Mr. Findlay[.]" *Id.* at 45:25–46:5. Microsoft defers to OpenAI.
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.	424. **Undisputed.** Microsoft defers to OpenAI.

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.	425. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI used Common Crawl for pre-training purposes. However, disputed as to the remainder of Paragraph 425 because the assertion is not supported by the cited evidence. Specifically, the cited report of Dr. Shan does not support the allegation that OpenAI acquired copies of books through "targeted 'scrapes' of particular websites." *See* ECF 1814-1 (S. Shan Decl. Ex. A) ¶¶ 111–12, 115–17. Specifically, the report discusses broad web crawlers like Common Crawl operated by third-parties, not "targeted 'scrapes' of particular websites" conducted by OpenAI. *See id.* ¶¶ 111–12, 115–17. Further, the remainder of the assertion is irrelevant because it has nothing to do with the works asserted by Book Plaintiffs. Microsoft defers to OpenAI.
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]	426. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI has created datasets with the names [redacted]" [redacted]" [redacted] and [redacted] However, disputed that the cited evidence supports the assertion that those datasets all contain data collected through a "targeted scrape," as defined by Plaintiffs to mean a "download of content available on a specific website." *See*

[redacted] Nelson Decl. Ex. 77 at 3; Nelson Decl. Ex. 351 at -872, -894.	Specifically, Mr. Sohl's testimony indicates that podcast data was not collected from "a specific domain." *See* C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 350:23-351:3. Microsoft defers to OpenAI.
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.*	427. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 427 are disputed because the facts are not supported by the cited evidence. Specifically, Book Plaintiffs mischaracterize web crawlers as programs that "download" website content and suggest that crawling and scraping are technologies particular to AI-related uses. OpenAI's expert witness Professor Feamster explained that web crawlers visit webpages and discover and follow links, whereas web scrapers extract specific content or data from webpages. *See* N. Feamster Opp. Decl. Ex. A ¶¶ 98–106, 107. Professor Feamster further explained that crawling and scraping technologies have numerous uses unrelated to AI and are widely used across industries. *See id.* ¶¶ 279–282. Microsoft defers to OpenAI.
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	428. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 428 are disputed because they are not supported by the cited evidence. Specifically, Book Plaintiffs' characterization of OpenAI's crawlers is not

augment search results displayed to users in ChatGPT. Nelson Decl. Ex. 72 at 2.	supported by the record evidence. There is no OpenAI user-agent called "SearchBot." OpenAI's crawlers include GPTBot, ChatGPT-User, and OAI-SearchBot, which serve distinct functions. *See* N. Feamster Opp. Decl. Ex. A ¶¶ 176–178; Gratz Decl. Ex. 190 (Overview of OpenAI Crawlers, OpenAI, https://developers.openai.com/api/docs/bots/). OpenAI further responds that the facts set forth in Paragraph 428 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Books Ex. 72 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Microsoft defers to OpenAI.
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.*	429. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 429 are disputed because they are not supported by the cited evidence. Specifically, Plaintiffs' characterization of OpenAI's user agents is not supported by the record evidence. There is no OpenAI user agent called "SearchBot"; the relevant user agent is "OAI-SearchBot." *See* Gratz Decl. Ex. 190 (Overview of OpenAI Crawlers,

	OpenAI, https://developers.openai.com/api/docs/bots/). Further, Book Plaintiffs' characterization that a user agent is "typically" how a bot appears to a website is inaccurate. Professor Feamster explained that, when properly configured, a bot identifies itself to websites through its user-agent string. *See* N. Feamster Opp. Decl. Ex. A ¶¶ 121–122. OpenAI also publicly provides information identifying its bots and publishes relevant egress IP ranges for its retrieval-related bots. *See id*. OpenAI further responds that the facts set forth in Paragraph 429 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Books Ex. 72 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Microsoft defers to OpenAI.
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	430. **Undisputed.** Microsoft defers to OpenAI.

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.	431. **Undisputed.** Microsoft defers to OpenAI.
432. Common Crawl operates the crawler known as "CCBot." *See* Nelson Decl. Ex. 72 at 4 (https://arxiv.org/pdf/2411.15091).	432. **Undisputed.** Microsoft defers to OpenAI.
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.	433. **Undisputed.** Microsoft defers to OpenAI.
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,	434. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that website operators may request that their pages not be crawled by specific crawlers by using "robots.txt." However, disputed that the cited evidence establishes that exclusion from Common Crawl requires a website owner to specifically add the "CCBot" user agent to a "disallow" list. Specifically, OpenAI expert Professor Feamster explained that a website operator may use the "*"

cannot modify that site's "disallow" list to prevent crawling of the pages containing the work. *See id.*	wildcard in a robots.txt file to disallow all crawlers, which would also prevent CCBot from crawling the website. See N. Feamster Opp. Decl. Ex. A ¶ 135. OpenAI further responds that the facts set forth in Paragraph 434 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Books Ex. 72 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Microsoft defers to OpenAI.
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.	435. **Undisputed** that the quotes appear in the cited support, but dispute any inferences drawn therefrom. Microsoft defers to OpenAI.
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	436. **Undisputed** that the quotes appear in the cited support, but dispute any inferences drawn therefrom. Microsoft defers to OpenAI.

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.	437. **Undisputed.** Defendants respond as follows: Undisputed that the quotes appear in the cited support, but dispute any inferences drawn therefrom.
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.").	438. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI used data from Common Crawl to train certain models. However, disputed that the cited evidence supports Book Plaintiffs' characterization that OpenAI obtained its third-party crawls "primarily" from Common Crawl or that OpenAI used Common Crawl content "extensively." Mr. Sohl agreed that Common Crawl was "one of the datasets that OpenAI used to train its models[,]" but did not testify that OpenAI primarily or extensively used that dataset. C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 283:13-17. And while Shan purports that OpenAI used Common Crawl "extensively," he does not qualify that use compared to other datasets. *See* ECF 1814-1 (S. Shan Decl. Ex. A) ¶¶ 112–17. Further disputed to the extent Book Plaintiffs' characterization that OpenAI "copied content from Common Crawl's crawl of the internet" suggests that

	OpenAI itself engaged in the underlying copying or crawling of third-party websites; Common Crawl conducted the web crawls and made the resulting data available. *See id.* at ¶¶ 112–113. Microsoft defers to OpenAI.
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).	439. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 351 contains the dataset names quoted in Paragraph 439, and that [redacted] as used within OpenAI may refer to Common Crawl. However, disputed to the extent Plaintiffs' statement implies that the [redacted] naming convention is uniform or that a dataset name alone establishes the provenance or contents of the dataset so named. The disputed proposition is not supported by the cited evidence. Specifically, Books Ex. 351 is a listing of dataset names and does not itself establish that all datasets containing [redacted] in their names were derived from data downloaded from Common Crawl, nor does it establish the provenance or contents of any dataset listed therein. *See* Books Ex. 351. Microsoft defers to OpenAI.
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.	440. **Undisputed.** Microsoft defers to OpenAI.

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.	441. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 441 are disputed because they are not supported by the cited evidence. Specifically, Paino did not acknowledge, state, or testify that OpenAI downloaded 56,000 books from b-ok. Rather, Paino testified that OpenAI accessed Common Crawl data, some of which had "B-OK" in the document domain. *See* C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 211:12–212:22. Plaintiffs' characterization of this testimony is materially misleading. Microsoft defers to OpenAI.
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.	442. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 442 are disputed because they are not supported by the cited evidence. Specifically, the cited testimony does not establish that Monaco "compiled" the list of 56,000 b-ok documents or portions of the text. Monaco's testimony states that he did not source documents from the b-ok domain; rather, Monaco testified the data was most likely sourced from Common Crawl. *See* C. Blythe Opp. Decl. Ex. 50 (V. Monaco 30(b)(1) Tr. (Jan. 28, 2026)) at 199:14-200:18. Microsoft defers to OpenAI.
443. Monaco testified that he compiled the list because he sought to identify instances of copyrighted book	443. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 443 are disputed because they are not

content in the training data. *See* Nelson Decl. Ex. 662 at 175:20–181:7, 199:14–200:18; Nelson Decl. Ex. 689.	supported by the cited evidence. Specifically, the cited testimony does not support Plaintiffs' characterization that Monaco "sought to identify instances of copyrighted book content in the training data." Monaco testified that he lacked "the expertise to make a judgment of whether or not something is copyright" and was instead "using that term generally to refer to what we were putting into or, you know, trying to put into the filter, more or less an engineer likely abusing a legal term in several ways." C. Blythe Opp. Decl. Ex. 50 (V. Monaco 30(b)(1) Tr.) at 179:15–180:13. Plaintiffs' characterization of Monaco's testimony as an acknowledgment that he sought to identify "copyrighted book content" is materially misleading, as Monaco expressly disclaimed the ability to make any copyright determination. Microsoft defers to OpenAI.
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.	444. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the FBI shut down Z-Library and arrested two individuals in connection with Z-Library's operation. However, disputed that Paino, Monaco, and other OpenAI employees "understood b-ok to be a mirror" of Z-Library, as that characterization is not supported by the cited evidence. *See* Paino testified that he "do[es] not know what B-OK means here" and did not "have a reason to definitively associate B-OK with Z-Library past that it is a Wikipedia page on Z-Library and it follows a paragraph where there's a quote from someone on the Authors Alliance about Z-

	library." C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 206:15–211:4. Monaco testified that he considered b-ok to be "in the same category as Z-Library," not that b-ok was a mirror of Z-Library, and testified that he understood the FBI seizure to relate to b-ok. C. Blythe Opp. Decl. Ex. 50 (V. Monaco 30(b)(1) Tr. (Jan. 28, 2026)) at 87:14–88:1. Books Ex. 798 likewise does not establish that b-ok was a mirror of Z-Library; the cited exchange reflects an employee's statement that a domain containing "b-ok" could have "Z-Library [] as the primary," while noting that "there were a bunch of mirrors." Further disputed that Plaintiffs accurately characterize the individuals arrested as Z-Library's "founders"; Books Ex. 733 describes them as operators of Z-Library. Microsoft defers to OpenAI.
445. OpenAI has also contributed financially to Common Crawl. In September 2023, OpenAI signed a contract to donate [redacted] to Common Crawl [redacted] 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").	445. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that in September 2023, OpenAI signed a contract that resulted in Common Crawl receiving a total of [redacted] in [redacted] annual payments of [redacted]. However, disputed to the extent Plaintiffs suggest that OpenAI's payment was a "donation" to Common Crawl to secure special influence or control over Common Crawl, as that characterization is not supported by the cited evidence. Specifically, Books Ex. 692 states that the [redacted] payment was a conditional grant to Common Crawl for the purpose of "enhanc[ing] Grantee's contributions to the

	thriving commons of open data that will drive innovation, research, and education in the 21st century as further described in Exhibit A to th[e] Agreement[.]" Moreover, the statement in Books Ex. 343 that "they're going to do whatever we want" was made in response to another employee's question of whether "it's better as a donation or commercial relationship," and does not support the inference that OpenAI exercised control over Common Crawl's operations. The Court should therefore disregard the disputed propositions in Paragraph 445 because the cited evidence is incorrect. Microsoft defers to OpenAI.
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.	446. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited exhibits contain the quoted statements. However, disputed to the extent Plaintiffs characterize OpenAI's payment as a "donation" to Common Crawl and suggest that the quoted purposes are the only purposes identified in the grant document, or that they were imposed as conditions on the grant. Books Ex. 692 states that the grant is "intended to be an unrestricted grant" and that the Grantee "will endeavor" to further the listed purposes, which also include "[i]ncreasing the coverage of currently under-represented languages in [Common Crawl's] crawl of the internet," "[l]abeling of content," and "[i]mprove security so that [Common Crawl] and its data are

	not used as a supply chain attack." The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.	447. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: This fact is disputed because Plaintiffs' characterization and inferences drawn therefrom are not corroborated by the cited evidence. Sohl testified that at "the time of this grant," "the increase in the size of Common Crawl's crawl in September of 2023 would not have significantly helped or hurt OpenAI" or "significantly influenced the total size of the datasets used to train [OpenAI's] models," and that he "do[es] not believe that the furtherance of [the] point" to "improve crawl strategy to seek out the highest-quality pages in each run" "would have significantly helped or hurt OpenAI." C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 282:9-284:2. Microsoft defers to OpenAI.

IV. OpenAI Trained Its LLMs on Datasets Containing Books

A. OpenAI Knew Books Were High-Quality Training Data

448. In June 2018, OpenAI published a paper titled "Improving Language Understanding by Generative Pre-	448. **Undisputed.** Microsoft defers to OpenAI.

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.	449. **Undisputed.** Microsoft defers to OpenAI.
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.	450. **Undisputed.** Microsoft defers to OpenAI.
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.	451. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 13 contains the quoted testimony. However, disputed that Ben Chess is currently "an OpenAI engineering manager." Chess is a former OpenAI employee who worked at OpenAI from 2019 to 2024, and whose title while at OpenAI was at one point engineering manager. *See* C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) 13:9-11, 18:1-15. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
452. Asked whether he understood that one type of high-quality data was books, Chess testified, "Seems	452. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Exhibit 13 contains the

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.	quoted testimony. However, disputed to the extent Plaintiffs imply that Chess affirmatively testified he understood that books were a type of high-quality data used for training OpenAI's models. Plaintiffs' characterization is inaccurate and incomplete. When asked whether he understood that "one type of high-quality data was books," Chess responded [redacted] *See* C. Blythe Opp. Decl. Ex. 12 (B. Chess 30(b)(1) Tr.) 41:6-42:1. Chess's testimony [redacted] Microsoft defers to OpenAI.
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.	453. **Disputed in Part.** OpenAI responds as follows: Undisputed that Books Ex. 50(a) contains the quoted testimony. However, disputed that Ryder was "[t]estifying as OpenAI's corporate designee" when he made the statement regarding people being "very preoccupied about this feature of books." OpenAI objected at the deposition that the question eliciting that testimony was outside the scope of Ryder's Rule 30(b)(6) designation. *See* C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr.) at 89:14-16. Ryder's response on that subject was not testimony offered in his capacity as OpenAI's corporate designee.

	Further, Plaintiffs' characterization is misleading because it omits qualifying language. Ryder actually testified, "*I think* in the early part of OpenAI, people were very preoccupied about this feature of books," whereas Plaintiffs omit the qualifier "I think." *See* C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr.) 89:17-90:7 (emphasis added). Plaintiffs' omission of this qualifier mischaracterizes Ryder's equivocal, personal impression as a definitive statement of institutional knowledge. Microsoft defers to OpenAI.
454. In November 2018, OpenAI wrote a document that set forth "Top level goals." Nelson Decl. Ex. 284.	454. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 454 are disputed because they are not supported by the cited evidence. Specifically, Plaintiffs' characterization of Exhibit 284 as a document that "set forth" OpenAI's "Top level goals" is not corroborated by the cited evidence. Although the document contains the phrase "Top level goals" in one place, it does not establish that the language identified as "Top level goals" constituted organizational goals of OpenAI. *See* Books Ex. 284 at OPCO_MDL_001687139. Microsoft defers to OpenAI.
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	455. **Disputed in Part.** OpenAI responds as follows: Undisputed that "build a large new text corpus for pretraining" was identified in Ex. 284 and that Dario Amodei commented it was a "very high priority (as well as

	likely to succeed)." *See* Books Ex. 284 at OPCO_MDL_001687140. However, disputed that "build a large new text corpus for pretraining" was the "first" of the "Top level goals" identified in the document. "Build a large new text corpus for pretraining" appears under a separate heading titled "Projects," not "Top level goals." *See* Books Ex. 284 at OPCO_MDL_001687139. Further disputed that the cited source establishes that Amodei was OpenAI's "Research Director" at the time of the document. In his deposition, [redacted] C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 17:15-21. Microsoft defers to OpenAI.
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.	456. **Disputed – Not Supported by Cited Evidence; Incomplete.** OpenAI responds as follows: This fact is disputed because Plaintiffs' statement is not supported by the cited testimony. Amodei testified in his individual capacity under Rule 30(b)(1), not as OpenAI's corporate designee. Pursuant to Fed. R. Evid. 106, Amodei testified, [redacted] C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 45:23–46:5. Plaintiffs' statement omits the qualifying phrase [redacted] and improperly characterizes Amodei's individual belief as establishing that scaling all aspects of large language model training was an OpenAI priority. The Court should therefore disregard the facts set forth in

	Paragraph 456 because they are not supported by the record. Microsoft defers to OpenAI.
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.	457. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited document is a draft of "GPT-3: Extreme Scaling of Transformer Language Models," and that the abstract states the authors "train[ed] a transformer language model with [redacted] parameters." Books Ex. 283 at OPCO_MDL_001687063. However, disputed as to Plaintiffs' characterization that "OpenAI" prepared the paper, as the cited evidence does not support that characterization. The Court should disregard the disputed proposition in Paragraph 457 because the cited evidence is incorrect. Microsoft defers to OpenAI.
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.	458. **Disputed in Part.** OpenAI responds as follows: Undisputed that the referenced document contains the quoted text. However, disputed as to the characterization that "OpenAI" added additional words from LibGen, and also as to the implication that the "combined datasets" included only the 6 billion word dataset used in Radford et al., 2019 and LibGen. The draft also states that "1B words from Wikipedia" were added. *See* Books Ex. 283 at OPCO_MDL_001687066.

	Microsoft defers to OpenAI.
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 20% of the time (despite being 40% of the data). This allowed for a good compromise between emphasizing the most useful data (webtext2) and having a lot of data.	459. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted language appears in the document. *See* Books Ex. 283 at OPCO_MDL_001687066. However, disputed that Book Plaintiffs' characterization that "OpenAI" conducted these actions with respect to the datasets is supported by the cited evidence as Ex. 283 does not establish that "OpenAI" took each of the listed actions. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
460. Notes from an August 2019 call between Microsoft and OpenAI employees record that OpenAI sought books from Microsoft because longer texts like books can improve a model's ability to handle "[l]ong term dependencies." Nelson Decl. Ex. 304 at -712 to -714_0001.	460. **Disputed – Not Supported by Cited Evidence.** Defendants respond as follows: The facts set forth in Paragraph 460 are disputed because they are not supported by the cited evidence. Specifically, Books Ex. 304 does not establish that OpenAI "sought books from Microsoft" or the purported reasons for doing so. The notes actually read: "David: Long term dependencies is tough. Do we have access to longer form things like books? (Read: do you have books databases) MSFT: Probably nothing specific. What about PDFs? One of the pieces of metadata

	is the length of the document." Books Ex. 304 at OPCO_MDL_01938714_0001. The cited evidence thus reflects a question posed during a call and not a request by OpenAI for books from Microsoft. Microsoft defers to OpenAI.
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.	461. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 305 contains the quoted language. However, disputed that the document "address[es] model performance per unit of computing power" as Plaintiffs characterize it, or that the quoted language reflects OpenAI's institutional position. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.*	462. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI published the paper announcing GPT-3 in 2020 and that the paper contains certain of the quoted language. However, disputed that the cited evidence supports Plaintiffs' characterization of the CommonCrawl dataset or of the role of Books1 and Books2. The cited evidence does not state that CommonCrawl was "of lower quality than 'more curated datasets'" as Plaintiffs assert; rather, it states that "*unfiltered or lightly filtered versions* of Common Crawl tend to have lower quality than more

	curated datasets." Books Ex. 162 at MSFT_AI_MDL_000101660. Moreover, the cited evidence does not support the assertion that OpenAI filtered CommonCrawl based on similarity to corpora "including Books1 and Books2." The paper states only that the authors filtered "based on similarity to a range of high-quality reference corpora," without defining those corpora as Books1 and Books2. Books Ex. 162 at MSFT_AI_MDL_000101660. The paper separately states that the authors "added several curated high-quality datasets, including . . . two internet-based books corpora (Books1 and Books2)," which is a distinct point from the filtering methodology. Books Ex. 162 at MSFT_AI_MDL_000101660. For those reasons, the disputed propositions are not supported by the cited evidence. Microsoft defers to OpenAI.
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.*	463. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited document contains the quoted language. However, disputed that the quoted language constitutes a statement by OpenAI reflecting OpenAI's corporate position. The cited evidence is a research paper, and Plaintiffs' characterization that "OpenAI further stated" misrepresents the nature of the document by attributing the paper's content to OpenAI as an institutional statement of its position. Books Ex. 162 at MSFT_AI_MDL_000101660.

	For that reason, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
464. The document that Dario Amodei at OpenAI sent to Saurabh Tiwary, Mikhail Parakhin, and Dee Templeton at Microsoft titled "Details, Results, and Benchmarking of OpenAI Large Models for MSFT" set out the same filtration approach, describing the construction of a CommonCrawl filter using high-quality datasets including books and the inclusion of those datasets in the training data mix. Nelson Decl. Ex. 273 at -524; Nelson Decl. Ex. 696.	464. **Disputed – Not Supported by Cited Evidence; Immaterial.** Defendants dispute the facts set forth in Paragraph 464 are disputed because they are not supported by the cited evidence. Specifically, the referenced document does not "set out the same filtration approach" as described in the preceding paragraph. Further, the cited document states that a "classifier filter" was constructed "based on a broad range of known high-quality datasets" and used "to remove low-quality data from CommonCrawl," but the document does not state that books were among the datasets used to create this filter. Books Ex. 273 at OPCO_MDL_001638524. Plaintiffs' assertion that the filter was constructed "using high-quality datasets including books" is therefore unsupported by the cited evidence. Microsoft further responds as follows: The facts set forth in Paragraph 464 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' summary judgment motion. Additionally, none of the testimony from Microsoft personnel about receiving the document cited as Books Ex. 273 demonstrates that Microsoft witnesses had personal knowledge of the sources of data in or contents of LibGen. *See* C. Blythe Opp. Decl. Ex. 51 (S. Tiwary 30(b)(1) Tr.) at

	30:8-24, 36:6-21 (with revision to 30:9 as detailed in errata); C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) at 68:8-25, 71:20-72:7; C. Blythe Opp. Decl. Ex. 54 (D. Templeton 30(b)(6) Tr. (Jan. 20, 2026)) at 285:13-286:26.
465. A table in that document rated LibGen [redacted] tokens and [redacted] of the training mix, and rated filtered CommonCrawl [redacted] and [redacted] of the training mix. Nelson Decl. Ex. 273.	465. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that Nelson Decl. Ex. 273 contains a table discussing LibGen and CommonCrawl. But Defendants dispute any implication in Paragraph 465 that the cited evidence supports that books themselves or the content of LibGen is [redacted] data for purposes of LLM training. While the cited evidence notes that the [redacted] of "Libgen" is [redacted], it is not clear what is meant by quality. Defendants also dispute Plaintiffs' characterization that the document "rated" these sources at particular [redacted], because the cited evidence does not establish what the [redacted] column means or how it was used.
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.*	466. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quotations. However, disputed that the cited evidence supports the characterization of the document as "discussing which datasets to include in a 2020 model training run" and the characterization of Ben Mann's comment as "not[ing] potential difficulties extracting the text." The disputed proposition is therefore not supported by the cited evidence.

	Microsoft defers to OpenAI.
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.	467. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed that Ilya Sutskever is currently OpenAI's Chief Scientist. Plaintiffs' characterization of Sutskever as "OpenAI's co-founder and Chief Scientist" implies he presently holds that title, which is not supported by the cited evidence. C. Blythe Opp. Decl. Ex. 55 (I. Sutskever 30(b)(1) Tr.) at 203:7-8. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.
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.*	468. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed to the extent Paragraph 468 implies that this exchange reflects OpenAI's opinion of book data quality. Specifically, the cited evidence reflects an exchange between two individuals and does not support the broader implication that it constitutes OpenAI's institutional opinion regarding book data quality. The disputed proposition is therefore not supported by the cited evidence. Microsoft defers to OpenAI.

469. An OpenAI document titled "GPT Wishlist" requested "more books (e.g. [redacted]), stating that books "are usually higher quality than websites" and that additional books "will help with [redacted] Nelson Decl. Ex. 269.	469. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the referenced document contains the quoted text. However, disputed that the document "requested" more books. The cited evidence does not support the characterization that the document constituted a "request" for more books; the meaning of the notation is subject to interpretation and appears to reflect a comment made during a brainstorming session rather than a formal request. For that reason, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	470. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed as to Plaintiffs' characterization of the document. For that reason, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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	471. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: Paragraph 471 is disputed because the cited evidence does not support Plaintiffs' assertion. When asked [redacted] Mann testified [redacted] C. Blythe Opp. Decl. Ex. 13

with long coherent text is high quality," with the first bullet underneath stating "watch libgen").	(B. Mann 30(b)(1) Tr.) at 263:18–264:12. Mann was then asked [redacted] and he responded [redacted]" *Id.* at 264:13-18. Mann [redacted] for those purposes, and the cited evidence does not support Plaintiffs' characterization. Microsoft defers to OpenAI.
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.	472. **Disputed in Part.** OpenAI responds as follows: Undisputed that Long Ouyang stated "i'm [sic] seeing gzip-filtering rejecting fairly high-quality books." Books Ex. 296 at OPCO_MDL_001897981. However, disputed as to Plaintiffs' characterization that "OpenAI researchers" made this statement. The cited evidence reflects a comment by a single researcher, Long Ouyang, describing his own observation, not a collective statement by multiple OpenAI researchers. Books Ex. 296 at OPCO_MDL_001897981. Microsoft defers to OpenAI.
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.	473. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed as to Plaintiffs' prefatory characterization that the Slack thread was "discussing the PDF portion of the Common Crawl web crawl," because Plaintiffs have not cited any evidence in support of that characterization.

	Microsoft defers to OpenAI.
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.	474. **Disputed in Part.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed to the extent Plaintiffs imply that Mr. Xu believed the Common Crawl contained many books. Plaintiffs' characterization is inaccurate because Xu testified [redacted]. C. Blythe Opp. Decl. Ex. 56 (T. Xu 30(b)(1) Tr.) at 57:4–59:23. Microsoft defers to OpenAI.
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.	475. **Undisputed.** Microsoft defers to OpenAI.
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.	476. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed as to Plaintiffs' characterization of Jain as "an OpenAI employee who worked on data issues." The cited

	evidence does not establish that Jain "worked on data issues." Microsoft defers to OpenAI.
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.*	477. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the accuracy of the quote. However, disputed as to Plaintiffs' characterization that Jain made the first statement "because" of the second. Although both statements appear in the cited exchange, the evidence does not establish the causal relationship Plaintiffs insert between them. For those reasons, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.*	478. **Undisputed.** Microsoft defers to OpenAI.
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.	479. **Undisputed.** Microsoft defers to OpenAI.
480. When Hoeschele asked whether there were "specific sources you're very excited about," Schulman	480. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows:

identified academic publications, books, and Substack, and stated that "books are generally higher quality than stuff on the internet." *Id.*	Undisputed that Books Ex. 279 contains the quoted statements attributed to Schulman. However, disputed that Plaintiffs' characterization that Schulman "identified" academic publications, books, and Substack as sources he was "very excited about" accurately reflects the cited exchange. Schulman initially responded that he "ha[d]n't thought about this much" and asked whether Hoeschele had "a list of prospects" before discussing those potential sources. *See* Books Ex. 279. Microsoft defers to OpenAI.
481. Schulman added that if OpenAI could improve "long context perf[ormance] in mid training" books would be valuable "for that reason." *Id.*	481. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 279 contains the quoted statement attributed to Schulman. However, disputed to the extent Plaintiffs characterize Schulman's statement as establishing that OpenAI viewed books as valuable for improving long-context performance. Schulman's statement was conditional—"whether we can improve long context perf in mid training"—and Ex. 279 does not establish that the condition was satisfied. Further disputed that the cited statement establishes that Schulman's use of "we" referred to OpenAI or what he meant by books being "valuable for that reason." *See* Books Ex. 279. Microsoft defers to OpenAI.
482. In November 2023, while discussing the performance of a competing model, Ryder stated that	482. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows:

"books" may explain that model's performance. Nelson Decl. Ex. 236 at -315.	Undisputed that the cited evidence reflects that, in November 2023, Ryder mentioned "books" while discussing a competing model's performance. However, disputed that Plaintiffs' characterization that Ryder "stated that 'books' may explain that model's performance" accurately reflects the cited evidence. Specifically, the cited evidence does not support the characterization that Ryder affirmatively "stated" that books "may explain" the competing model's performance; rather, the cited exchange reflects only that Ryder mentioned "books" in the context of discussing that model's performance. *See* Books Ex. 236 at OPCO_MDL_000157315. Microsoft defers to OpenAI.
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.*	483. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited evidence reflects the quoted language from Radford's messages. However, disputed as to Plaintiffs' characterization implied by the insertion of the bracketed word "[as]" into Radford's quoted statement. Plaintiffs' insertion of "[as]" alters the meaning of Radford's original language, and the cited evidence does not support the sentence as Plaintiffs have presented it. *See* Books Ex. 236. Microsoft defers to OpenAI.
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.	484. **Disputed in Part.** OpenAI responds as follows: Undisputed as to the accuracy of the quoted language attributed to Brockman. However, disputed as to

books, university coursework/notes, etc) will become the most valuable." Nelson Decl. Ex. 356 at -585.	Plaintiffs' characterization of that statement. Plaintiffs omitted the portion immediately preceding the quoted language, in which Brockman wrote "but we can guess that as we get better at synthetic data approaches . . ." The omitted prefatory language establishes that Brockman's statement was speculative in nature, not a definitive assertion that such data sources "will become the most valuable." Plaintiffs' selective quotation thus presents a misleading characterization of the cited evidence. *See* Books Ex. 356 at OPCO_MDL_004776585. Microsoft defers to OpenAI.
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.	485. **Disputed in Part.** OpenAI responds as follows: Undisputed as to the accuracy of the quoted language attributed to Brockman. However, disputed as to Plaintiffs' characterization that Brockman "expected" a data amplifier. The cited evidence reflects that Brockman discussed the "data amplifier" concept as one potential scenario, presented in contrast to a hypothetical world where a model could learn "entirely from synthetic data." Plaintiffs' use of the word "expected" mischaracterizes the nature of Brockman's discussion by presenting a contingent, comparative scenario as a definitive expectation. *See* Books Ex. 356 at OPCO_MDL_004776580. Microsoft defers to OpenAI.
486. The understanding that books were valuable training data was widely shared in the LLM industry,	486. **Disputed.** OpenAI responds as follows: Disputed as mischaracterizing and overstating the evidence.

including among OpenAI's competitors Anthropic and Google, and among third-party researchers. Zhao Decl. Ex. A at ¶¶ 46–72.	The Zhao report discusses individual documents/publications from Google and Anthropic indicating that at least some individual researchers might have thought books were valuable training data. The evidence does not support the conclusion that this was an "understanding" that was "widely shared in the LLM industry." Microsoft defers to OpenAI.
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.	487. **Disputed – Not Supported by Admissible Evidence.** Defendants respond as follows: The facts set forth in Paragraph 487 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Disputed as mischaracterizing the document. It states "Hard to quantify, but overall we believe that books represent very high quality pretraining data." This is speculation, not a set statement as Plaintiffs imply. *See* C. Callison-Burch Opp. Decl. Ex. B ¶¶ 52-55. Microsoft defers to OpenAI.
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.	488. **Disputed – Not Supported by Admissible Evidence; Not Supported by Cited Evidence.** Defendants respond as follows: The facts set forth in Paragraph 488 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. The facts set forth in Paragraph 488 are disputed because the

	cited evidence is incorrect. Specifically, the document is titled "Data Hypotheses" and does not purport to convey any findings. Rather, the introduction from the drafter says "This document tries to list various concrete hypotheses I've had about data. I don't believe any of them with very high probability. In many cases, I think they're more likely than not to be wrong. The goal is just to try to list/name/articulate some things which might be true about data." They conflate "The Coherent, Long-Context Hypothesis" with "The Good Writing Hypothesis" which states: "Models just need good writing. The kind of content an editor would approve of. This teaches them how to write well, rather than low quality internet speak." See Books Ex. 741 at ANT_BARTZ_000003433. Microsoft defers to OpenAI.
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.	489. **Disputed – Not Supported by Admissible Evidence; Not Supported by Cited Evidence.** Defendants respond as follows: The facts set forth in Paragraph 489 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. The facts set forth in Paragraph 489 are disputed because the cited evidence is incorrect. Specifically, this is one employee's thoughts, as set forth in the "Data Hypotheses" document, and he said "they're more likely than not to be wrong."

	Microsoft defers to OpenAI.
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.	490. **Undisputed.** Microsoft defers to OpenAI.
491. When describing how it would effectuate this strategy, Anthropic stated that data quality would be a "key contributor." *Id.*	491. **Undisputed.** Microsoft defers to OpenAI.
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.*	492. **Undisputed.** Microsoft defers to OpenAI.
493. Books were listed first on Anthropic's list of sources of additional "High-Quality Data." Nelson Decl. Ex. 741.	493. **Undisputed.** Microsoft defers to OpenAI.
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.	494. **Disputed in Part – Not Supported by Admissible Evidence; Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed as to the quote. However, disputed because there is nothing on the face of the document to demonstrate this was stated by "members of the strategic finance team" (*see* C. Callison-Burch Opp. Decl. Ex. A ¶¶ 52-55) so this proposition is not supported by the cited evidence.

	Further, the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the evidence cited is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability under Fed. R. of Ev. 803(18), and therefore may not be considered in support of summary judgment. Microsoft defers to OpenAI.

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.	495. **Disputed – Not Supported by Cited Evidence.** Defendants respond as follows: Disputed that "OpenAI's contemporaneous documents establish that it regarded books as among the highest-quality training data available and pursued them for that reason" because Plaintiffs do not explain what timeframe "contemporaneous" refers to. OpenAI otherwise incorporates Defendants' responses from Section IV.A., *supra*, as Plaintiffs rely on it for fact 495. Further disputed that Defendants' experts do not contend that books failed to improve model performance. Rather, OpenAI's experts stated that books were not necessarily required as a technical matter. Plaintiffs' own cited support undermines this characterization.

	Specifically, the cited paragraphs of Dr. Callison-Burch's Opening Report state that using only books was not helpful for creating generalist models. Books Ex. 751 (C. Callison-Burch Opening Rep.) ¶¶ 331–32. Dr. Callison-Burch's Opening Report further states that "[n]o individual work or set of works is required to achieve a diverse distribution of data." *Id.* ¶ 328. Similarly, the cited support for Dr. Lafferty states that books are not uniquely necessary to general LLM capability. Books Ex. 727 (J. Lafferty Rebuttal Rep.) ¶ 65. For those reasons, the disputed proposition is not supported by the cited evidence.
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.	496. **Disputed in Part – Incomplete.** Defendants respond as follows: Defendants do not dispute that certain experts offered opinions that Defendants did not need the Asserted Works to create general-purpose large language models and that no individual work was indispensable to those models' capabilities. However, Defendants dispute that the cited experts ever stated or used the words "not technically required." The relevant opinions of the cited experts are provided below. Defendants also dispute that "no category of works including books was indispensable," as that characterization is not supported by the cited evidence. Dr. Lafferty stated "no individual text or category of texts within a training corpus of this scale is necessary to the model's capability." Pursuant to Fed. R. Evid. 106, Dr. Lafferty's full opinion is that "[t]raining on text from a

	limited range of sources would limit the ability of the model to generalize and to generate novel responses. To avoid this, developers aggregate material from multiple categories of documents into the training material, including long-form prose, technical writing, and informal conversation, among many others. The resulting model can then generate language that varies in tone, vocabulary, and structure. The greater the diversity in the training corpus, the more robustly the model learns patterns that are characteristic of language as a whole rather than patterns specific to any particular source or category of text. For the same reason, no individual text or category of texts within a training corpus of this scale is necessary to the model's capability; the learned patterns are properties of the aggregate distribution, and 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. Because the statistical patterns the model learns are distributed redundantly across millions of sources spanning trillions of tokens, no individual work or collection of works is singularly responsible for those patterns." Books Ex. 753 ¶¶ 86-87. Dr. Tucker opined that no one piece of data is essential to training a general-purpose LLM. Pursuant to Fed. R. Evid. 106, Dr. Tucker's full opinion is: "No 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. For this reason, the incremental value of adding any one piece of data, such as an individual news article, to a training dataset set is close to zero." Books Ex. 754 ¶ 98. Dr. Callison-Burch opined that LLMs' success "in no way depends on the inclusion of any specific document or group of documents in the training set." Books Ex. 751 ¶ 344. But, pursuant to Fed. R. Evid. 106, OpenAI responds to note that Plaintiffs ignore his broader opinion that "in a large-scale pre-training operation, what matters is the composition of the training corpus as a whole, primarily because the purpose of pre-training is to identify broad patterns that manifest in numerous different examples and contexts. No individual work or set of works is required to achieve a diverse distribution of data; indeed, any individual source, particularly in a given domain, is entirely fungible." *Id.* ¶ 328. Further, Dr. Callison-Burch opined that "[p]articularly critical here is diversity: to create an LLM capable of comprehending and responding to the full scope of possible user queries and use-cases, OpenAI needed a dataset that included near-universal coverage across every possible dimension, including subject matter, sources, forms of language (formal, colloquial), styles of writing, etc." *Id.* ¶ 329. Dr. Callison-Burch also opined that "[t]o the extent any of the plaintiffs' works were part of that massive and diverse corpus of natural language text used to pretrain the disputed OpenAI LLMs and products, it was incidental

	to, and solely to serve, this new and different purpose." *Id.* ¶ 327.
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.	497. **Undisputed – Incomplete.** OpenAI does not dispute that Dr. Callison-Burch made the statements in Paragraph 328 in his report. But, pursuant to Fed. R. Evid. 106, OpenAI incorporates Defendants' response regarding their experts' testimony cited in response to Paragraph 496. Microsoft defers to OpenAI.
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).	498. **Undisputed – Incomplete.** OpenAI does not dispute that Dr. Callison-Burch made the statements in Paragraph 11(c) in his report. But, pursuant to Fed. R. Evid. 106, OpenAI incorporates Defendants' response regarding their experts' testimony cited in response to Paragraph 496. Microsoft defers to OpenAI.
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.	499. **Undisputed – Incomplete.** OpenAI does not dispute that Dr. Callison-Burch made the statements in Paragraphs 343-44 in his report. But, pursuant to Fed. R. Evid. 106, OpenAI incorporates Defendants' response regarding their experts' testimony cited in response to Paragraph 496. Microsoft defers to OpenAI.
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	500. **Undisputed – Incomplete.** OpenAI does not dispute that Dr. Callison-Burch made the statements in Paragraph 20(a) in his reply report. But, pursuant to Fed.

that "no individual work was necessary for inclusion in the corpora." Nelson Decl. Ex. 756, ¶ 20(a).	R. Evid. 106, OpenAI incorporates Defendants' response regarding their experts' testimony cited in response to Paragraph 496. Microsoft defers to OpenAI.
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.	501. **Undisputed – Incomplete.** OpenAI does not dispute that Dr. Callison-Burch's testimony included the referenced quotation. But, pursuant to Fed. R. Evid. 106, OpenAI incorporates Defendants' response regarding their experts' testimony cited in response to Paragraph 496. Microsoft defers to OpenAI.
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.	502. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statements in Paragraph 502 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
503. Lafferty states that "no individual work or subset of works is responsible for conferring" an LLM's capacity to generalize. *Id.* ¶ 21.	503. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 503 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496.

	OpenAI defers to Microsoft's response.
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.	504. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statements in Paragraph 504 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
505. Lafferty further opines that "news and book content are not uniquely necessary to LLM capability." Nelson Decl. Ex. 727 at ¶ 5.	505. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 505 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
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.	506. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 506 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
507. He states that "books may be useful as a training signal since they offer long-form coherence,	507. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr.

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.	Lafferty made the statements in Paragraph 507 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
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.*	508. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 508 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
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.	509. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 509 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
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.	510. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty made the statement in Paragraph 510 in his report. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates

	its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. OpenAI defers to Microsoft's response.
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.	511. **Undisputed – Incomplete.** Pursuant to Fed. R. Evid. 106, Microsoft incorporates its response to Paragraph 496 regarding Dr. Tucker's opinions. OpenAI defers to Microsoft's response.
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.*	512. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Tucker's report included those statements in Paragraph 512 from the depositions of Mr. Usdan and Mr. Hoeschele. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response to Paragraph 496 regarding Dr. Tucker's opinions. OpenAI defers to Microsoft's response.
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	513. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Tucker's report contains the statements quoted in Paragraph 513. But, pursuant to Fed. R. Evid. 106, Dr. Tucker's full opinion is: "The 211 Asserted Works are a miniscule share of the datasets used to train the GPT models." Books Ex. 754 ¶ 100. Microsoft also incorporates

[Microsoft's] models." *Id.* ¶ 133 (second alteration in original).	its response to Paragraph 496 regarding Dr. Tucker's opinions. OpenAI defers to Microsoft's response.
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.	514. **Undisputed – Incomplete.** Defendants do not dispute that Mr. Ryder gave the quoted testimony. But pursuant to Fed. R. Evid. 106, Mr. Ryder also testified that, although there were "intuitions about the kind of text we find very high quality" and "many investigations," "all of those investigations, for the purposes of pretraining, always led back to the most important factor by far is getting rid of spam, getting rid of SEO, getting rid of obscene material," and that "the primary goal was to maximize the amount of human-written words, spoken words, thoughts." C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr.) 33:1–8, 174:23-175:5. When asked whether books are "high-quality long-context training data," Mr. Ryder explained that [redacted] C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr.) 89:14–90:10.
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.	515. **Undisputed.** Microsoft defers to OpenAI.

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.	516. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Mr. Usdan gave the quoted testimony. But, pursuant to Fed. R. Evid. 106, Mr. Usdan also explained that LLMs are trained on "trillions and trillions of words," so making a measurable difference requires "very, very, very large data sets, millions and millions of books," and a smaller set of books is "like a drop in the ocean." Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) at 79:7–18. He further testified that it is "scientifically . . . extremely hard to judge the impact of a data set in a text model." Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) at 80:18–25; *see also* ECF 1707 (J. Usdan Decl.) ¶¶ 10–11, 22, 24(a). OpenAI defers to Microsoft's response.

C. OpenAI Used LibGen to Train GPT-3 and GPT-3.5

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.	517. **Undisputed.** Microsoft defers to OpenAI.
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.	518. **Undisputed.** Microsoft defers to OpenAI.

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.	519. **Disputed in Part – Not Supported by Cited Evidence**. Defendants respond as follows: Undisputed that LibGen1 and LibGen2 were included in certain datasets, such as the dataset for [redacted]. However, disputed to the extent Plaintiffs' characterization suggests LibGen1 and LibGen2 were used to train any model as of April 2019, as that inference is contradicted by Plaintiffs' own assertions in SUF Nos. 517 and 518 that the LibGen datasets had not even been created until on or after that date. Further disputed that the cited evidence supports the assertion that the model demonstrated to Bill Gates and other senior Microsoft executives in April 2019 was "an early version of GPT-3." The disputed propositions are not supported by the cited evidence, and Books Ex. 62(a) is incorrect. Disputed as to the claim that OpenAI "demonstrated" the purported "GPT-3" model "to Bill Gates and other senior Microsoft executives." Bill Gates was no longer a Microsoft executive in April 2019 and did not have decision-making authority over Microsoft's investment in OpenAI. *See* C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 65:6-67:13, 109:3-110:2. Umesh Madan joined the in-person meeting by telephone to take notes, but no other Microsoft personnel were in attendance at the referenced April 2019 meeting. *See* Books Ex. 188 at -958; C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 15:24-16:1. Further, Mr. Gates and Mr. Madan both testified that they did not have discussions with other "Microsoft senior executives"

	concerning the datasets used to train the model that OpenAI demonstrated. C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 138:2-139:11; C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 52:20-25. Further disputed as none of the testimony from "Microsoft executives" demonstrates that they had personal knowledge of the sources of data in or contents of LibGen. C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:7-22; *see id.* at 52:20-25; C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 115:6-116:14, 119:2-14, 123:12-124:9.
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.	520. **Undisputed.** Microsoft defers to OpenAI.
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.	521. **Undisputed.** Microsoft defers to OpenAI.
522. In May 2021, [redacted] was a [redacted] token mixture used in [OpenAI's] largest runs," containing [redacted] tokens from LibGen1 and [redacted] tokens from LibGen2. Nelson Decl. Ex. 316.	522. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 522 are disputed because they are not supported by the cited materials. Disputed that the cited evidence supports the asserted token counts for LibGen1 and LibGen2 as no token counts are included in Books Ex. 316. Microsoft defers to OpenAI.

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.	523. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted language appears at the cited Bates number. However, disputed to the extent Plaintiffs rely on the phrase "the equivalent of [redacted] books" out of context to suggest that [redacted] actual books were acquired from LibGen or another source for training GPT-3. The statement on its face refers to "the equivalent of" a number of books based on an extrapolation from the size of books in the training mix — not to the actual number of books used. Indeed, in the same document, Jain states that "libgen1 is 110k docs, libgen2 is 450k docs" — referring to "docs," not "books," totaling approximately 560,000 documents, not [redacted] books. Books Ex. 349 at OPCO_SDNY_1636095. The cited evidence therefore does not support Plaintiffs' characterization. Microsoft defers to OpenAI.
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.	524. **Undisputed.** Microsoft defers to OpenAI.
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.	525. **Undisputed.** Microsoft defers to OpenAI.

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.	526. **Undisputed.** Microsoft defers to OpenAI.
527. Testifying as OpenAI's corporate designee, Ian Sohl confirmed that [redacted] was the data mix used for GPT-3.5. Nelson Decl. Ex. 57(a) at 448:16–20.	527. **Undisputed.** Microsoft defers to OpenAI.
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.	528. **Undisputed.** Microsoft defers to OpenAI.
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.	529. **Undisputed.** Microsoft defers to OpenAI.
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.	530. **Undisputed.** Microsoft defers to OpenAI.
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.	531. **Undisputed.** Microsoft defers to OpenAI.

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.	532. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 532 are disputed because they are unsupported by the cited evidence. Specifically, the cited testimony and document relate only to a GitHub pull request referencing The Pile; they do not show that the GitHub pull request was approved, do not contain any reference to Books3 or Bibliotik, and do not discuss model training at all, let alone model training on The Pile. C. Blythe Opp. Decl. Ex. 43 (V. Misra 30(b)(1) Tr.) 36:10–38:10; Books Ex. 277 at OPCO_MDL_001667186. Microsoft defers to OpenAI.

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.	533. **Disputed – Improper Legal Argument; Not Supported by Cited Evidence.** OpenAI responds as follows: Paragraph 533 is disputed, and the Court should disregard this Paragraph because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. In particular, Plaintiffs' characterization that OpenAI "copied" books into its training datasets calls for a legal conclusion. Further, the facts set forth in Paragraph 533 are disputed because the cited evidence is incorrect. The Shan report cited by

	plaintiffs does not state that GPT-4o mini was not trained on datasets derived from Common Crawl. *See* ECF 1814-1 (S. Shan Decl. Ex. A) ¶¶ 117–119, 134–138; C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 450:5-12; C. Blythe Opp. Decl. Ex. 57 (I. Sohl Ex. 420). Microsoft defers to OpenAI.
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.	534. **Disputed in Part.** OpenAI responds as follows: Undisputed that GPT-4o mini is a performant and functional large language model and undisputed that Plaintiffs have not asserted that GPT-4o mini trained on any data containing the Asserted Works. However, disputed that OpenAI trained GPT-4o mini [redacted]. *See* C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 450:5-12; Ex. 57 (Sohl Ex. 420). Microsoft defers to OpenAI.

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.	535. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that training large language models involves analyzing large corpora of text. However, disputed that the training process "requires" creating "multiple copies" of the data at each stage of the training pipeline, because the cited evidence does not support that proposition. Specifically, the

	cited evidence does not establish that training a large language model "requires" making "multiple copies" of the data at each stage; rather, as OpenAI's corporate designee testified, "model training itself is not making any -- anything past ephemeral copies in memory. There's no copying." C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 302:14-16. He further explained that the model code reads data from storage into a computer's memory so it can be "read" or processed. *Id.* at 302:17-303:1; *see also* C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 535 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	536. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI converts raw text files in preparing training datasets. However, disputed that this process creates a "copy," as the conversion involves modifications to the original data and results in different data than before. *See* C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 536 because the assertion constitutes

	improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	537. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI filters training datasets. However, disputed that filtering creates an additional "copy." The characterization that filtering datasets creates an additional "copy" is disputed because filtering involves modifications to the original data and results in different data than before, not a duplicate of the original. *See* C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 537 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	538. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI samples training datasets. However, disputed that sampling creates an additional "copy." The characterization that sampling datasets creates an additional "copy" is disputed because sampling involves modifications to the original data and results in different data than before, not a duplicate of the original. *See* C. Callison-Burch Opp. Decl.

	Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 538 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	539. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI cleans training datasets. However, disputed that cleaning creates an additional "copy." The characterization that cleaning datasets creates an additional "copy" is disputed because the process involves modifications to the original data and results in different data than before, not a duplicate of the original. *See* C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 539 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	540. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI tokenizes training data. However, disputed that tokenization creates an additional "copy." The

	characterization that tokenizing data creates an additional "copy" is disputed because tokenization involves converting text files to tokens, which modifies the original data and results in different data than before, not a duplicate of the original. C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 540 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	541. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI shards tokenized data. However, disputed that sharding creates an additional "copy." Sharding involves modifications to the original data and results in different data than before. C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 541 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.

542. OpenAI creates an additional copy when it caches the data. Shan Decl. Ex. B ¶ 41.	542. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI caches data. However, disputed that caching creates an additional "copy." Caching involves modifications to the original data and results in different data than before. C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 542 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	543. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI loads the dataset into the model training environment. However, disputed that loading the dataset into the model training environment creates an additional "copy." Loading data into the model training environment involves modifications to the original data and results in different data than before. C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 543 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response.

	Microsoft defers to OpenAI.
544. Copying repeats each time OpenAI passes the dataset through the model during a training epoch. Shan Decl. Ex. B ¶¶ 41–42.	544. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that OpenAI passes the dataset through the model during a training epoch. However, disputed that "copying repeats" each time the dataset is passed through the model during a training epoch. Passing data through the model during a training epoch involves modifications to the original data and results in different data than before. C. Callison-Burch Opp. Decl. Ex. B ¶¶ 59, 62; C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (July 16, 2026)) at 510:6-512:1. Further, the Court should disregard the disputed portion of Paragraph 544 because the assertion constitutes improper attorney argument and/or is conclusory as to what constitutes a "copy." It thus requires no response. Microsoft defers to OpenAI.
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.	545. **Undisputed** that the cited document contains metadata from June 2022 and lists multiple file paths to OpenAI Azure container locations that have the string "libgen" in the file path, including multiple file paths that also have the string "pretokenized" in the file path as well. Microsoft defers to OpenAI.

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.	546. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed that the cited paragraph of Dr. Shan's report supports the statement, because it does not contain citations to any evidence in support of his assertions. Rather, Dr. Shan simply states *ipse dixit* that "Weng created numerous copies of the datasets" without any citations to the record. ECF 1814-1 (S. Shan Decl. Ex. A) ¶ 70. Microsoft defers to OpenAI.
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.	547. **Disputed.** OpenAI responds as follows: This fact is disputed because it mischaracterizes the cited evidence. Paino testified that during training, the training source code "grabs some bytes" from data from persistent storage (the "BLOB store"), that the retrieved "bytes exist in memory," that he "wouldn't really call [the bytes that exist in memory during training] a real copy," and that it "would be a little bit silly" to attempt to count how many times bytes are temporarily loaded to memory during training. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 302:11–303:6. Microsoft defers to OpenAI.

G. The Accused Models Reproduce Class Plaintiffs' Works to End Users

548. The Accused Models have the capability to, and in fact do, reproduce portions of the Asserted Works to	548. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows:

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.	Defendants respond that this fact is disputed because Paragraph 548 mischaracterizes how the Accused Models operate to the extent that the term "reproduce" implies that any of the Accused Models store copies of the Asserted Works. LLM outputs are generated using next-token prediction, and instances of the models at issue "reproduc[ing] portions of the Asserted Works" are extremely rare. *See* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶ 33; ECF 1723 (J. Lafferty Decl.) ¶ 35; ECF 1707 (J. Usdan Decl.) ¶¶ 19-20. There is no evidence that the Accused Models and products have ever regurgitated *any* portion of the vast majority of Asserted Works. Rather, the cited evidence shows, at most, that any alleged regurgitation, of even short passages, of Asserted Works is an extremely rare and unintended side-effect of the training process rather than the intended operation of the model. Defendants further dispute that any assertions regarding "requests for summaries" are supported by the cited evidence because the cited evidence does not discuss or show any requests for summaries. Moreover, the cited evidence also does not cover all Accused Models; there are no "regurgitations" or "reproduc[tions]" from GPT-3, GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o, and GPT-4o mini identified anywhere in the cited evidence.

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.	549. **Disputed in Part – Not Supported by Cited Evidence.** Defendants incorporate their response to 548. Disputed to the extent this statement asserts that Dr. Shan "documented actual reproduction" of all Asserted Works because the cited evidence does not support it. As to ChatGPT, Dr. Shan identified only 33 limited instances of alleged "regurgitation" from just 7 of the Asserted Works, and for Copilot, Dr. Shan identified only 24 limited instances of alleged "regurgitation" of 30 or more near-verbatim words from only 10 Asserted Works (and in both instances most come from a single Asserted Work (Game of Thrones)). *See* ECF 1814-1 (S. Shan Decl. Ex. A) at ¶¶ 41, 154, 156; T. Berg-Kirkpatrick Opp. Decl. ¶¶ 54-55; ECF 1723 (J. Lafferty Decl.) p. 12 n. 13 (Shan found limited matches of 10 Asserted Works across 24 conversations); *id.* at Table 1 (Shan's matches account for 0.0003% of the 8.2 million Copilot dataset). Further, pursuant to Fed. R. Evid. 106, Defendants note that the "reproduction[s]" Dr. Shan identified examples are highly concentrated in a small number of works and authors. T. Berg-Kirkpatrick Opp. Decl. ¶ 55. Defendants further dispute any implication that "reproduc[es]" Asserted Works in everyday conversations with users. OpenAI's expert Dr. Berg-Kirkpatrick analyzed 95 million user-visible ChatGPT outputs and found that 25-word verbatim excerpts or near-verbatim regurgitations of Class Plaintiffs' content appeared in 14 outputs, which

	corresponds to 0.00007%. ECF 1750 (T. Berg-Kirkpatrick Decl. ¶ 30). Based on this analysis, Dr. Berg-Kirkpatrick concluded that ChatGPT outputs containing verbatim or near-verbatim portions of Plaintiffs' Asserted Works are so rare that the vast majority of ChatGPT users will never encounter one. Id. ¶ 33.
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.	550. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as to one of the 33 alleged "regurgitations" identified by Dr. Shan. OpenAI's expert Dr. Berg-Kirkpatrick reviewed the excerpts of the logs Dr. Shan identified and was unable to locate one of the claimed "reproductions." Berg-Kirkpatrick Opp. Decl. ¶ 54. Dr. Shan later acknowledged this mistake. ECF 1814-6 (Shan Decl. Ex. F) ¶ 54. Further, pursuant to Fed. R. Evid. 106, the "reproduction[s]" Dr. Shan identified examples are highly concentrated in a small number of works and authors. T. Berg-Kirkpatrick Opp. Decl. ¶ 55. OpenAI further disputes any implication that "reproduc[es]" Asserted Works in everyday conversations with users. OpenAI's expert Dr. Berg-Kirkpatrick analyzed 95 million user-visible ChatGPT outputs and found that 25-word verbatim excerpts or near-verbatim regurgitations of Class Plaintiffs' content appeared in 14 outputs, which corresponds to 0.00007%. ECF 1750 (T. Berg-Kirkpatrick Decl. ¶ 30). Based on this analysis, Dr. Berg-Kirkpatrick

	concluded that ChatGPT outputs containing verbatim or near-verbatim portions of Plaintiffs' Asserted Works are so rare that the vast majority of ChatGPT users will never encounter one. Id. ¶ 33. Microsoft defers to OpenAI.
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.	551. **Disputed in Part.** Microsoft responds as follows: Microsoft does not dispute that Plaintiffs' expert Dr. Shawn Shan analyzed the output logs Microsoft produced in this litigation or that Dr. Shan claimed to identify the reproductions identified in Paragraph 551. But Microsoft disputes that the 24 "chat responses" were user-facing outputs of Bing Chat (Copilot). Dr. Shan did not separate text that users saw from Microsoft's internal search-and-retrieval ("grounding") data. He testified that he matched the displayed text and the internal material "together." C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) 291:10-14. Microsoft also disputes Shan's findings. Dr. Lafferty's corrected analysis yielded a much smaller count. He ran a 16-word sequence match over the Books data from Dr. Shan's analysis, separated user-facing [redacted] from internal [redacted] (grounding data), and removed matches to user prompts. That analysis found matches in only 4 conversations in the user-facing [redacted] field and 8 in the [redacted] grounding field. ECF 1723 (J. Lafferty Decl.) ¶¶ 371, 378 & Table 5 (Rebuttal).

	Microsoft also incorporates its response to Paragraph 548. OpenAI defers to Microsoft's response.
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.	552. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the produced output log samples are only a fraction of all output logs that exist. However, disputed that "[t]he output log samples OpenAI produced generally do not encompass the earlier periods when regurgitation would have been more prevalent in earlier periods[.]" Nothing in the cited paragraphs from Dr. Shan's expert report sets forth any analysis or evidence supporting this assertion; instead, Dr. Shan states "I understand that the output log samples OpenAI produced in this case generally do not encompass earlier periods of time when regurgitation would have been more prevalent" without citation to any evidence in the record. ECF 1814-1 (S. Shan Decl. Ex. A) ¶ 151. It is therefore a mere *ipse dixit*. Moreover, Dr. Berg-Kirkpatrick showed that deduplication measures were in place for GPT-3 (fuzzy deduplication) and that regurgitation rates are vanishingly low across all measured periods. ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 25, 33. The "lower bound" framing is an expert characterization dependent on Shan's methodology, which Krishnan has challenged for lacking error-rate analysis. *See* S. Krishnan Opp. Decl. (Books) Ex. A ¶ 64 (noting that Shan inspected less than 3% of his claimed matches, provides no error-rate assessment, and

	his lack of validation results in false positives); *see also* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 23, 25-26; ECF 1723 (J. Lafferty Decl.) ¶¶ 26–28. Further, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	553. **Disputed in Part.** OpenAI responds as follows: Undisputed that an OpenAI employee wrote the quoted statement in the cited document. However, disputed as to the context and characterization of the quoted statement because "regurgitation" does not appear in the cited document, the cited document does not state when the 78 million-log sample was collected, and the cited document does not connect the quoted statement to any "regurgitation studies." Therefore, the cited evidence does not support the assertion that 78 million logs are a "little light" for evaluating regurgitation. Moreover, Dr. Berg-Kirkpatrick explained in his expert report that a 78M sample size is sufficient for a regurgitation analysis. *See* T. Berg-Kirkpatrick Opp. Decl. ¶ 21. Microsoft defers to OpenAI.
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.	554. **Disputed in Part – Expert Opinion Subject to Rule 702.** OpenAI responds as follows: Undisputed that the produced logs contained redactions to protect user privacy. However, disputed that the redactions "imped[ed] identification of infringing outputs" as the sole cited paragraph from Dr. Shan's Opening Report provides

	only a single example of a redaction, and Dr. Berg-Kirkpatrick explained that there are multiple methods to account for redactions in the produced logs while conducting a regurgitation analysis. *See* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 14-15. Dr. Berg-Kirkpatrick conducted a near-verbatim regurgitation analysis designed to identify overlapping texts between model outputs and Plaintiffs' AWs despite word differences like those produced by redacted text— Plaintiffs' experts could have conducted a similar analysis. Dr. Berg-Kirkpatrick's opening report results demonstrated that no instances of 25-gram regurgitation of Plaintiffs' AWs were obstructed by either one or two redactions. ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 14-16, 22. Even more basic techniques were available to Plaintiffs (e.g., treating redacted tokens as wildcards). Further, the evidence cited in support of this proposition is subject to a motion to exclude expert testimony pursuant to Fed. R. Evid. 702. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Microsoft defers to OpenAI.
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	555. **Disputed – Improper Legal Argument; Immaterial; Not Supported by Cited Evidence.** Defendants respond as follows: Defendants respond that this fact is disputed because Paragraph 555 misstates the technical operation of LLMs. Defendants incorporate their response to Paragraph 548.

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.	Further, the Court should disregard this Paragraph because the assertion constitutes improper attorney argument and is conclusory. It thus requires no response. Specifically, disputed to the extent the statement suggests that a plot summary is an unlawful "reproduction" under copyright law. A plot-faithful summary discussing named characters, key plot points, and story conclusions is a description of unprotectable elements, and is not a "reproduction" of any of the Asserted Works. Characterizing summaries as "reproduction" is an argumentative legal conclusion embedded in a factual statement. Further, the facts stated in Paragraph 555 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. The facts set forth in Paragraph 555 are disputed because the cited evidence does not support the asserted proposition. Also disputed because the assertion that requests for summaries are "[o]f the kind users make every day" finds no support in the cited evidence. The underlying data points (word counts, prompt types) are also disputed because they are based on the same criticized methodology discussed in Fact 552. And Dr. Berg-Kirkpatrick analyzed Dr. Shan's summary outputs and found that they contained errors such as misidentified central events and incorrect plot details, and therefore were not "plot-faithful descriptions." *See* T. Berg-Kirkpatrick Opp. Decl. ¶ 41; S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 48-50.

	Microsoft further responds that none of the cited evidence supports that Copilot will produce multi-paragraph, plot-faithful descriptions averaging approximately 400 words and reaching a maximum of 2,353 words. Dr. Shan did not run this test on Copilot.
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.	556. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Disputed to the extent that it suggests Dr. Shan was able to generate summary outputs containing verbatim passages of thirty or more words from multiple Asserted Works and from multiple Accused Models. The cited evidence does not identify the "named Plaintiffs' books" or the models that generated the six summary outputs. But as Dr. Berg-Kirkpatrick explained in his reply report, one of the six summary outputs was generated from GPT-4-base, an unaligned base model that was neither publicly available nor intended for public use. And the remaining five summary outputs all came from one book, A Game of Thrones, and one model, GPT-4o, and all five involved the same underlying quote that is readily available on numerous public websites and in the Amazon Kindle sample of A Game of Thrones. *See* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 59, 70-71; T. Berg-Kirkpatrick Opp. Decl. ¶ 42. Further, the disputed proposition in Paragraph 556 is immaterial because it does not affect the outcome of the suit under the governing law. Further, the disputed proposition is not supported by the cited evidence.

	Microsoft defers to OpenAI.
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.").	557. **Disputed in Part.** OpenAI responds as follows: Disputed because the statement does not contain the full quote from Mr. Monaco, which does not reference training data. Pursuant to Fed. R. Evid. 106, OpenAI notes that the cited transcript shows that Mr. Monaco stated that "the objective of Project Giraffe was to develop techniques to further limit regurgitation in a particular context." Books Ex. 38(a) (V. Monaco Jan. 27, 2026 30(b)(6) Tr.) at 35:15–18. Microsoft defers to OpenAI.
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.	558. **Disputed in Part.** OpenAI responds as follows: Undisputed that Mr. Monaco's cited testimony includes the quoted text. However, disputed to the extent it implies that such regurgitation was a common phenomenon and there were no measures deployed to address it. Pursuant to Fed. R. Evid. 106, Monaco further testified that "going back to around the time ChatGPT launched, it was general knowledge in the field, and we were aware that models can memorize and regurgitate data. And we wanted to prevent that." Books Ex. 38(b) (V. Monaco 30(b)(6) Tr. (Apr. 8, 2026)) at 212:2–5. Therefore, Mr. Monaco's testimony establishes that memorization was a known, limited phenomenon.

	Microsoft defers to OpenAI.
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.	559. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted text appears in the cited document. Disputed as to the characterization of the cited document and quoted text. The assertion that OpenAI "described memorization not as an accident but as a product of the training objective itself" materially mischaracterizes Books Ex. 298 and what it says about combating regurgitation. "Repeat data -> memorization -> regurgitation" (Books Ex. 298 at OPCO_MDL_001900469) means that memorization arises from data duplication, not the training objective. Books Ex. 298 further describes "Masked Training" as a concrete mitigation and concludes that OpenAI "[c]an prevent regurgitation in LMs by clever masking" and the approach is "mostly working." *Id.* at OPCO_MDL_001900480. And Books Ex. 298 also reports that anti-regurgitation measures reduced the probability of reproducing a specific code snippet from 2.6% to approximately 3×10^{-6} (a reduction of approximately 870,000×). *Id.* at OPCO_MDL_001900477. The statement in Books Ex. 298 that "this is the first time we are trying to make LLMs bad at something" merely reflects that anti-regurgitation was a novel engineering challenge to be solved; it does not suggest that LLMs are designed to memorize. The characterizations of Ex. 298 are therefore not supported by the evidence, and the quotes come solely

	from the speaker notes of a single engineer. Similarly, the proposition that memorization was "positively correlated with [OpenAI's] own evaluations of model quality" is not supported by Ex. 298. Nothing in the cited materials supports that assertion. *See id.* at OPCO_MDL_001900469, -470, -477. Further, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	560. **Disputed in Part.** OpenAI responds as follows: Undisputed that the quoted text appears in the cited document without the ellipses and with additional text that the quote omits. Disputed to the extent it suggests that OpenAI trains neural networks to only memorize datasets. The quoted portions of Ex. 274, which is a Slack thread, come from a single engineer who is providing a simplified answer to a teaching question. The portions of Ex. 274 that are omitted from the quoted text provide important context to the statement: the engineer is discussing a model trained on the same data multiple times, and the quoted excerpts omit the main clause from the engineer's statements, which says generalization emerges from training. The engineer also explains in the same response that "catastrophic forgetting" lets the model read a book again, "building deeper general understanding without memorizing too much." And the engineer explains that OpenAI studies these issues "because we want to train the best generalizable neural networks." Books Ex. 274 at

	OPCO_MDL_001644366. Therefore, this statement is disputed because it misleadingly omits important context from within and around the quoted text. *See also* C. Callison-Burch Opp. Decl. Ex. B ¶ 134. Microsoft defers to OpenAI.
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.	561. **Disputed in Part.** OpenAI responds as follows: Undisputed that the quoted text appears in the cited document, except the quote is missing a tilde "~" character before the term "everything." The actual quote in the cited document (Books Ex. 286) has the text "memorize ~everything" where the tilde denotes "approximately." Disputed to the extent Plaintiffs suggest that Books Ex. 286 concerns training objectives or possibilities concerning the Accused Models. Indeed, the referenced quote discusses assumptions about "future big models" and plans for future mitigation tools. *See* Books Ex. 286 at OPCO_MDL_001697816; C. Callison-Burch Opp. Decl. Ex. B ¶ 117. Microsoft defers to OpenAI.
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.	562. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quoted text appears in the cited paragraph of Dr. Shan's Rebuttal Report. However, disputed as unsupported by any cited source material, which makes it impossible to verify the accuracy of the quote, the completeness of the quote, or the context in

	which the quote arose. The only evidence cited for this assertion is a quote from Dr. Shan's Rebuttal Report, which does not identify the speaker, the date, or the context for the quote. Also disputed to the extent it characterizes memorization as a "problem," which is based on Dr. Shan's opinion but is unsupported by any source document. Further, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	563. **Disputed in Part.** OpenAI responds as follows: Undisputed that the document is titled "Update on Preventing Regurgitation" and that the quoted text appears in the document. Disputed as to the date attribution of March 2022, as the cited document does not contain any date, and further disputed to the extent it suggests that all models within the GPT-3.5 family of models "can easily reproduce training data verbatim." Moreover, the context for the quoted text in Books Ex. 292 is source code and personal information, not any text from any Asserted Works. Microsoft defers to OpenAI.
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.	564. **Disputed.** OpenAI responds as follows: This fact is disputed because the claim that OpenAI released these models "without any anti-regurgitation protections" is factually inaccurate and is rebutted by primary evidence in the record. The GPT-3 paper (Books Ex. 680) describes

	performing "fuzzy deduplication at the document level, within and across datasets" — a pre-training-level measure specifically designed to reduce memorization. The internal anti-regurgitation presentation (Books Ex. 298) itself identifies "Repeat data -> memorization -> regurgitation" and describes deduplication as a key mitigation pathway. OpenAI employee Ben Mann testified [redacted] C. Blythe Opp. Decl. Ex. 13 (B. Mann 30(b)(1) Tr.) 423:1–21. The statement therefore ignores pre-training-level anti-regurgitation protections (deduplication, data filtering, regularization), rendering the claim factually inaccurate. The statement that OpenAI "recogniz[ed] internally that regurgitation had not been solved" is also disputed. While Ex. 298 states that regurgitation was not "100% solved," the statement implies that OpenAI acted with deliberate indifference when the evidence shows that OpenAI actively made ongoing engineering investments to address regurgitation (masked training, deduplication improvements, etc.). *See* Books Ex. 680 at 3; Books Ex. 298 at OPCO_MDL_001900469, -470, -477. Microsoft defers to OpenAI.
565. OpenAI did not deploy the initial version of Giraffe filter to ChatGPT until a few weeks after ChatGPT's	565. **Undisputed.**

November 30, 2022 launch. Nelson Decl. Ex. 38(b) at 212:6–16.	Microsoft defers to OpenAI.
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.	566. **Undisputed.** Microsoft defers to OpenAI.
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.	567. **Disputed in Part.** OpenAI responds as follows: Disputed that the cited evidence shows that all of Plaintiffs' Works are present throughout the files used in or generated as part of Project Giraffe. The cited paragraph from Dr. Shan's Opening Report differs from the assertion (it uses a disjunctive "or" instead of "and") and cites only two files, which does not show that Plaintiffs' works are "present throughout" all Project Giraffe files. The cited paragraph also fails to explain how Dr. Shan found the 46 works, or what form they take in the files, leaving it unclear and unsupported what it means to "reflect" those 46 "Plaintiffs' Works." Microsoft defers to OpenAI.
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.	568. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited evidence contains references to locations in Vinnie Monaco's personal Azure container containing the label "search" or "book-search." Disputed

	that the cited evidence supports Dr. Shan's characterization of the Azure containers or how the data therein was used. Dr. Shan does not point to any OpenAI documents or testimony to explain what data is stored in these Azure containers or how they were used. Instead, Dr. Shan makes assumptions based only on the fact that these locations hit on specific search terms. Further disputed to the extent that the statement implies that these Azure locations contain copies of any of the Asserted Works or portions thereof, as the cited paragraph from Dr. Shan's Opening Report only states, without analysis, that OpenAI's anti-regurgitation team "located" 96 of Plaintiffs' works without any explanation of what was searched or whether any result contained text from the books themselves (as opposed to text from blog posts, reviews, or other sources that are different from the books themselves). Further, the disputed proposition is not supported by the cited evidence. Microsoft defers to OpenAI.
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.	569. **Disputed in Part.** OpenAI responds as follows: To the extent this statement suggests that all of the Asserted Works were "regurgitated" by all of the Accused Models, the evidence does not support that assertion. Moreover, the cited paragraphs from Dr. Shan's Opening Report include only conclusory characterizations of two internal OpenAI files. The cited paragraphs do not identify which model was tested, what prompts were used, or how much text was purportedly reproduced. Taken at face value,

	the files at most implicate only 8 of the Asserted Works, but provide no analysis of the two identified files. Microsoft defers to OpenAI.
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.	570. **Disputed – Not Supported by Cited Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 570 are disputed because they are not supported by the cited evidence. Specifically, the proposition that the Accused Models "memorized at least portions of Plaintiffs' Works" and "stored those portions inside the model" is a disputed expert assertion and inference, not a statement of fact supported by the cited evidence. As Dr. Callison-Burch has explained, LLMs do not store specific text from training data, are not databases, and generate text through prediction rather than through retrieval of stored text. *See* Books Ex. 751 (C. Callison-Burch Opening Rep.) ¶ 11; C. Callison-Burch Opp. Decl. Ex. A ¶¶ 96–98, 141. Dr. Callison-Burch further states that "Dr. Shan does not contest my point that models are not databases" and that LLMs "are designed to store patterns, rather than storing information." C. Callison-Burch Opp. Decl. Ex. C ¶ 116. Model parameters are numerical weights, not stored text; no verbatim text exists or is "stored" within the model's weight matrices. Moreover, across 95 million ChatGPT messages in the 20M Merits Sample, verbatim 25-word excerpts from all Plaintiffs' asserted works appeared only 31 times—less than 0.000033% of outputs—and the Named Book Plaintiffs' rate was less than 0.00001%. *See*

	ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 23, 25-. Corrected books regurgitation in user-visible Copilot output was 4 messages, representing 0.000003% of the 8.2M dataset and 0.0000004% using the 1B denominator. *See* ECF 1723 (J. Lafferty Decl.) Table 5. Microsoft defers to OpenAI.

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.	571. **Disputed – Not Supported by Cited Evidence.** Defendants dispute that "OpenAI and Microsoft exchanged copies of the LibGen data with each other." The cited evidence does not support that claim. The cited evidence does not mention any instance in which Microsoft shared LibGen data with OpenAI. *See* ECF 1814-1 (S. Shan Decl. Ex. A) ¶ 72. Further, the documents cited in the referenced paragraph of Shan's report show that, with regard to transfer of training data to Microsoft, OpenAI only provided Microsoft with a subset of its GPT-3 training data in a tokenized format. *See* C. Blythe Opp. Decl. Ex. 39 (B. Norick 30(b)(1) Tr.) at 47:11; C. Blythe Opp. Decl. Ex. 78 (B. Gaffney 30(b)(6) Tr.) at 148:5–15; C. Blythe Opp. Decl. Ex. 60 (Song 30(b)(1) Tr.) at 83:5–13; C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 237:11–13. OpenAI later only provided Common Crawl, WebText2, and English Wikipedia. *See* Books Ex. 275. As explained by Callison-Burch in response to Shan's contentions in the referenced paragraph, "documents produced in this

	litigation suggest that of the five datasets used to train GPT-3 . . . OpenAI only transferred three to Microsoft, none of which contained LibGen data." C. Callison-Burch Opp. Decl. Ex. B ¶ 62(d). Indeed, as OpenAI's corporate designee, Gaffney testified that OpenAI transferred only a "subset" of training data to Microsoft. C. Blythe Opp. Decl. Ex. 78 (Gaffney 30(b)(6) Tr.) at 148:2-6.
572. OpenAI also distributed LibGen data to third-party contractors. Shan Decl. Ex. A ¶ 72; Nelson Decl. Ex. 42 at 111:4–112:16.	572. **Disputed – Not Supported by Cited Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 572 are disputed because they are not supported by the record. Specifically, OpenAI did not distribute LibGen content to third-party contractors. Said contractors were permitted only to view excerpts of books one at a time through OpenAI's website interface. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 127:7-128:18; C. Blythe Opp. Decl. Ex. 59 (L. Ouyang 30(b)(1) Tr.) at 140:23-141:4. Microsoft defers to OpenAI.

A. <u>OpenAI Provided Microsoft Access to GPT-3 Training Data</u>

573. Under a September 2020 amendment to the 2019 OpenAI-Microsoft agreement, OpenAI agreed to provide [redacted]	573. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that the quoted statements in Paragraph 573 are contained in the September 2020 amendment to the 2019 JDCA. But Defendants dispute that the amendment required OpenAI to provide the "full set" of GPT-3 training data. The record

[redacted] Nelson Decl. Ex. 158 at -967.	shows that OpenAI only provided Microsoft with a subset of its GPT-3 training data in a tokenized format. *See* C. Blythe Opp. Decl. Ex. 39 (B. Norick 30(b)(1) Tr.) at 47:11 (explaining that OpenAI transferred "[a] fraction of pretraining data"); C. Blythe Opp. Decl. Ex. 78 (B. Gaffney 30(b)(6) Tr.) at 148:5–15 (explaining that OpenAI granted Microsoft access to a subset of GPT-3 training data); C. Blythe Opp. Decl. Ex. 60 (Song 30(b)(1) Tr.) at 83:5–13 ("One-third of the data was in the tokenized format, was transferred for the only purpose of evaluation … , and we had to delete that data after the days that we got assigned to us for the evaluation, which I believe was 90 days, or something like that. So that's the whole extent when it comes to the data exchange."); C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 237:11–13 ("[T]he training datasets were tokenized, so not in natural language for a human to look at."). Further, OpenAI agreed to provide Microsoft [redacted] *See* Books Ex. 158 at MSFT_AI_MDL_000019967. Exhibit E to the referenced agreement expressly [redacted] *Id.*

574. OpenAI's corporate witness Benton Gaffney agreed that this provision entitled [redacted] Nelson Decl. Ex. 18 at 145:14–146:4.	574. **Disputed in Part.** Defendants do not dispute that Mr. Gaffney agreed to the statement in Paragraph 574 at his deposition. But Defendants dispute any implication in Paragraph 574 that Mr. Gaffney agreed that the provision referenced in Paragraph 573 granted Microsoft [redacted]. Microsoft received [redacted] *See* Books Ex. 158 at MSFT_AI_MDL_000019964 (GPT-3 license [redacted] *see also* C. Blythe Opp. Decl. Ex. 78 (Gaffney 30(b)(6) Tr.) at 148:5-15 (explaining that OpenAI granted Microsoft temporary access to a subset of GPT-3 training data).
575. [redacted]	575. **Undisputed – Incomplete.** Defendants do not dispute the fact stated in Paragraph 575 but, pursuant to Fed. R. Evid. 106, further respond that OpenAI only provided Microsoft with a subset of its GPT-3 training data in a tokenized format. *See* C. Blythe Opp. Decl. Ex. 39 (B. Norick 30(b)(1) Tr.) at 47:11 (explaining that OpenAI transferred "[a] fraction of pretraining data"); C. Blythe Opp. Decl. Ex. 78 (B. Gaffney 30(b)(6) Tr.) at 148:5–15 (explaining that OpenAI granted Microsoft temporary access to a subset of GPT-3 training data) C. Blythe Opp.

	Decl. Ex. 60 (Song 30(b)(1) Tr.) at 83:5–13 ("One-third of the data was in the tokenized format, was transferred for the only purpose of evaluation … ."); C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) 237:11–13 ("[T]he training datasets were tokenized, so not in natural language for a human to look at.").
576. [redacted]	576. **Undisputed.**
577. Microsoft accessed the GPT-3 training data in January 2021, verified that the data appeared to be correct, and later removed "these assets from the repo." Nelson Decl. Ex. 346; Shan Decl. Ex. A ¶ 72(d).	577. **Disputed in Part – Immaterial.** Microsoft does not dispute that in January 2021, a limited number of Microsoft engineers accessed a portion of OpenAI's GPT-3 training data only for the purpose of evaluating OpenAI's GPT-3 model, and no longer had access once the evaluations were completed. C. Blythe Opp. Decl. Ex. 60 (X. Song 30(b)(1) Tr.) 82:25–83:13. But Microsoft disputes the assertion that Microsoft "verified that the data appeared to be correct." The cited evidence states only that Microsoft engineers "worked together to *accept the GPT-3 model source code* from OpenAI." Books Ex. 346 at OPCO_MDL_004635421. Microsoft's engineers have testified that the goal of the evaluations was "to make sure when OpenAI agreed to transfer these GPT-3 assets to Microsoft, we, as a recipient of the assets, we actually received what they were claimed to be. We are receiving functional things, functional assets, so

	we had to run an evaluation on top of the assets we received to make sure they work as expected." C. Blythe Opp. Decl. Ex. 60 (X. Song 30(b)(1) Tr.) 84:23–85:5; C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) 237:11–13 ("[T]he training datasets were tokenized, so not in natural language for a human to look at."). Microsoft further responds that the facts set forth in Paragraph 577 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. Additionally, none of the testimony from Microsoft's engineers responsible for evaluating the GPT-3 model demonstrates that they had personal knowledge of the sources of data in or contents of LibGen. *See* C. Blythe Opp. Decl. Ex. 60 (X. Song 30(b)(1) Tr.) 83:15–21 ("I have no knowledge of the source, where it came from."); *id.* at 90:10–24 ("Q. Are you familiar with Library Genesis, Mr. Song?" "A. No, not at all." "Q. Have you ever heard of Library Genesis before today?" "A. If I did, it did not register with me."); C. Blythe Opp. Decl. Ex. 51 (S. Tiwary 30(b)(1) Tr.) at 36:6–21 (declining knowledge about OpenAI datasets); C. Blythe Opp. Decl. Ex. 39 (B. Norick 30(b)(1) Tr.) 20:6–10 ("I didn't look at the data used to train GPT-3, no.").
578. OpenAI also shared portions of the GPT-3 training data with Microsoft on other occasions. Nelson	578. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that OpenAI shared portions of GPT-3 training data with

Decl. Ex. 41 at 46:15–47:11; Nelson Decl. Ex. 275 at -591; Shan Decl. Ex. A ¶ 72(d).	Microsoft in November 2021. But Defendants dispute any implication in Paragraph 578 that OpenAI transferred the GPT-3 training datasets to Microsoft at large; only four engineers received temporary access to a portion of the GPT-3 training data under strict use permissions and confidentiality obligations. *See* Books Ex. 275 at OPCO_MDL_001653590-91 [redacted] Defendants also dispute that OpenAI shared portions of the GPT-3 training data with Microsoft "on other occasions" after November 2021 because the cited evidence does not support that assertion. Defendants further respond that the facts set forth in Paragraph 578 are immaterial because these facts do not affect the outcome of the suit under the governing law. Plaintiffs cite no evidence that the subset of GPT-3 training data that was transferred to Microsoft in November 2021 contained data from any of the Asserted Works. No data that OpenAI obtained from LibGen was transferred to Microsoft in November 2021. *See* Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021).
579. In November 2021, Microsoft shared access to GPT-3 training datasets [redacted]	579. **Disputed – Not Supported by Cited Evidence.** Defendants dispute that Microsoft shared

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.	access to GPT-3 training datasets [redacted] The cited evidence describes that *OpenAI*, not Microsoft, shared access to certain portions of the GPT-3 training datasets with limited Microsoft personnel for [redacted] Books Ex. 1 (K. Allison 30(b)(1) Tr.) 82:4–20; Books Ex. 780 at OPCO_MDL_002072434. Defendants also dispute any implication that Microsoft transferred the full Bing Index to OpenAI. ECF 1814-1 (S. Shan Decl. Ex. A) ¶ 127. Microsoft transferred only *portions* of the Bing Index to OpenAI. C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) 101:19–102:16 (testifying only a "subset" of the Bing Index was "sent to OpenAI").
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.	580. **Disputed – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** Defendants respond as follows: The facts set forth in Paragraph 580 are disputed because they are not supported by the cited evidence. Specifically, the cited evidence does not support that OpenAI "sought a copy of the 'Bing Index'" from "the outset of its relationship with Microsoft" without limitation. While Exs. 163 and 159 reference the possibility of sharing some amount of data from the Bing Index, they do not reflect a request from OpenAI. *See* Books Ex. 163 at MSFT_AI_MDL_000118609; Books Ex. 159 at MSFT_AI_MDL_000046571.

	Further, Defendants respond that the facts set forth in Paragraph 580 are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited constitutes inadmissible hearsay, and the Microsoft emails on which Plaintiffs rely are objectionable for lack of foundation and personal knowledge, as they cannot supply OpenAI's intent.
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.	581. **Disputed in Part – Incomplete.** Defendants do not dispute that Microsoft and OpenAI entered into a [redacted] on November 6, 2019, or that pursuant to that agreement, Microsoft agreed to make available to OpenAI certain data from Microsoft's Bing Index. Defendants dispute that this agreement described the data as only [redacted] Pursuant to Fed. R. Evid. 106, the agreement provided: [redacted] Books Ex. 161 at MSFT_AI_MDL_000097042.
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]	582. **Disputed in Part.** Defendants do not dispute that the parties amended the statement of work on March 25, 2020, but Defendants respond that the original November 2019 statement of work (Books Ex. 161) included

[redacted] Nelson Decl. Ex. 178.	the same provision providing that OpenAI could use the Bing Index data for computational analysis. *See* Books Ex. 178 at MSFT_AI_MDL_000533953.
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.	583. **Undisputed.**
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.	584. **Disputed.** Microsoft disputes that Microsoft's "Bing Index crawler" (Bingbot) can access paywalled or so-called "copyrighted" content that web publishers restrict Bingbot from accessing. Microsoft's Bingbot is a compliant crawler that comports with responsible crawling practices; if Bingbot is able to crawl paywalled or so-called "copyrighted" content, it is because the website owner has not restricted Bingbot from doing so. ECF 1705 (P. Gaggara Decl.) ¶¶ 10-23. Further, Microsoft disputes Paragraph 584 to the extent that it relies on the testimony of Jordan Usdan. Pursuant to Fed. R. Evid. 106, Mr. Usdan stated he was "not a subject matter expert on how Microsoft handles paywalled content" and suggested that plaintiffs "depose people in Bing to get a deeper understanding." C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) 138:11–140:4. OpenAI defers to Microsoft's response.

585. Microsoft raised copyright concerns over providing OpenAI with the Bing Index. 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.	585. **Disputed – Immaterial; Improper Legal Argument.** Defendants respond that the facts set forth in Paragraph 585 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Specifically, "concerns" are not facts and are immaterial to the issues before the court. Defendants also dispute any implication in Paragraph 585 that Microsoft's provision of portions of Bing Index data to OpenAI violated copyright law. The Court should disregard this Paragraph because the assertion constitutes improper attorney argument and/or is conclusory. OpenAI further disputes Plaintiffs' assertion that OpenAI "continued to press for data" through 2021 and 2022; Ms. Gaggara's testimony in Books Ex. 19(a) discusses the provision of Bing Index data pursuant to the Statements of Work. C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) 111:1–119:9.
586. Microsoft was "aware of the legal considerations for the Books1 [i.e., LibGen-1] and Books2 [i.e., LibGen-2] datasets." Nelson Decl. Ex. 291.	586. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Microsoft responds that the facts set forth in Paragraph 586 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is hearsay and lacks foundation; it is a statement by an OpenAI employee on what "Bing" is

	allegedly aware of. Books Ex. 291 at OAI_MDL_001833041. Books Ex. 291 is an OpenAI Slack thread that includes only OpenAI employees. Microsoft responds that the facts set forth in Paragraph 586 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. OpenAI defers to Microsoft's response.
587. Following negotiations between OpenAI's Chief Technology Officer, Mira Murati, and Microsoft, the parties agreed [redacted] *Id.*	587. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial.** Undisputed that Books Ex. 291 contains the quoted language. However, Microsoft disputes facts set forth in Paragraph 587 as they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is hearsay and lacks foundation as to Microsoft; it is a statement by an OpenAI employee as to actions being taken by OpenAI. Defendants further respond that the facts set forth in Paragraph 587 are immaterial and therefore disputed because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See* ECF 1705

	(P. Gaggara Decl.) ¶¶ 20–23 ("Microsoft takes measures to avoid including unauthorized copyrighted content in the Bing Index"); Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the LibGen datasets). Thus, these facts do not affect the outcome of the suit under the governing law.
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.	588. **Disputed – Incomplete; Not Supported by Cited Evidence.** Paragraph 588 mischaracterizes the cited evidence. Books Exhibit 779 does not support the facts set forth in Paragraph 588. Pursuant to Fed. R. Evid. 106, in response to a summary of Ms. Murati's statement described in Paragraph 587, Mr. McGrew stated: "Is the idea that we're giving data to them and they are giving data to us, [redacted] followed by "Or is it that we're only going to train [redacted] Books Ex. 779 at OPCO_MDL_001936752. The cited evidence does not support that Mr. McGrew described the so-called "arrangement" in the language of the quoted text. Mr. McGrew posed a question; he did not offer a conclusion that either question was correct.
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."	589. **Disputed – Incomplete; Not Supported by Cited Evidence; Immaterial.** Microsoft disputes the facts set forth in Paragraph 589. Microsoft did not condition the Bing Index transfer on OpenAI's provision of training data. Karmel Allison testified: [redacted]

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.	[redacted]. Books Ex. 1 (K. Allison 30(b)(1) Tr.) at 68:20–23, 72:1–8; Books Ex. 173 at MSFT_AI_MDL_000254440. Microsoft also disputes that Ms. Allison [redacted] because it is not supported by the cited evidence. Microsoft employee Misha Bilenko stated in an email to Ms. Allison: [redacted] Books Ex. 173 at MSFT_AI_MDL_000254441. Karmel Allison then responded, "[redacted] *Id*. at MSFT_AI_MDL_000254440. When asked whether she was [redacted] Books Ex. 1 (K. Allison 30(b)(1) Tr.) at 69:2–23. Microsoft further disputes the claim that Ms. Allison used the phrase "horse trading" to describe the purported "arrangement" concerning Microsoft's transfer of Bing Index data and OpenAI's provision of GPT-3 training data. Ms. Allison's testimony referred to an email presented at her deposition that stated: "there would have to be horse-

	trading around *the GPUs*." GPUs ("graphic processing units") are computing hardware components, not data. C. Blythe Opp. Decl. Ex. 83 at MSFT_AI_MDL_000096529. Ms. Allison also testified that [redacted] Books Ex. 1 (K. Allison 30(b)(1) Tr.) 70:25–72:8. Microsoft also responds that the facts set forth in Paragraph 589 are immaterial and therefore disputed because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See* ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23 ("Microsoft takes measures to avoid including unauthorized copyrighted content in the Bing Index"); Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the LibGen datasets). Thus, these facts do not affect the outcome of the suit under the governing law. OpenAI joins Microsoft's response.
590. Katherine Mayer testified that OpenAI [redacted] because it was [redacted]	590. **Disputed in Part – Incomplete; Immaterial.** Defendants do not dispute that Books Ex. 826 contains the quoted testimony. Defendants state that Paragraph 590 leaves out relevant context from the witness pursuant to Fed. R. Evid. 106. *See* C. Blythe Opp. Decl. Ex.

[redacted]. Nelson Decl. Ex. 826 at 163:13–25.	64 (K. Mayer 30(b)(1) Tr.) at 164:5–12 ([redacted]). Defendants also respond that the facts set forth in Paragraph 590 are immaterial and therefore disputed because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See* ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23 ("Microsoft takes measures to avoid including unauthorized copyrighted content in the Bing Index"); Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the LibGen datasets). Thus, these facts do not affect the outcome of the suit under the governing law.
591. OpenAI [redacted] Nelson Decl. Ex. 781 at -435 (emphasis added). [redacted]" Nelson Decl. Ex. 173.	591. **Disputed in Part – Incomplete; Immaterial.** Defendants do not dispute that Books Ex. 781 contains the following statement by Katie Mayer on November 15, 2021: "[redacted] Books Ex. 781 at OPCO_MDL_002072435

	(emphasis in original). But, pursuant to Fed. R. Evid. 106, Ms. Mayer testified that by [redacted] she meant [redacted] C. Blythe Opp. Decl. Ex. 64 (K. Mayer 30(b)(1) Tr.) at 160:25–161:4. Defendants also do not dispute that Nelson Decl. Ex. 173 contains a statement from Parakhin that states, [redacted] Books Ex. 173 at MSFT_AI_MDL_000254439. Defendants dispute any implication in Paragraph 591 that Microsoft "sold" the Bing Index in exchange for GPT-3 training data. *See* Books Ex. 1 (K. Allison 30(b)(1) Tr.) at 68:20–23, 72:1–8 ([redacted] and [redacted].). Defendants also respond that the facts stated in Paragraph 591 are immaterial and therefore disputed because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See* ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23 ("Microsoft takes measures to avoid including unauthorized copyrighted content in the Bing Index"); Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the

	LibGen datasets). Thus, these facts do not affect the outcome of the suit under the governing law.
592. [redacted]	592. **Disputed – Incomplete; Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Parakhin agreed with the quoted statement during his deposition. Microsoft does dispute any implication in Paragraph 592 that what Microsoft may have "told" OpenAI reflects an action that Microsoft actually took. Pursuant to Fed. R. Evid. 106, Karmel Allison stated, [redacted] Books Ex. 1 (K. Allison 30(b)(1) Tr.) at 69:20–23; Books Ex. 173 at MSFT_AI_MDL_000254440. Microsoft further responds that the facts stated in Paragraph 592 are immaterial and therefore disputed because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See* ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23 ("Microsoft takes measures to avoid including unauthorized copyrighted content in the Bing Index"); Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the LibGen datasets). Thus, these facts do not affect the outcome of the suit under the governing law. OpenAI joins Microsoft's response.

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.	593. **Disputed in Part – Immaterial.** Defendants do not dispute that Microsoft's provision of portions of the Bing Index to OpenAI was called "Project Taxi." Defendants dispute that Microsoft provided "full copies of the Bing Index to OpenAI." The Bing Index data that Microsoft provided to OpenAI excluded webpages applying the NOARCHIVE and/or NOCACHE meta tag directives, and Microsoft removed any web content subject to DMCA takedown protocols. C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan 27, 2026)) 230:22-231:8 ("We removed anything which was 'noarchive/nocache' from them. We removed anything which was DMCA removed from them."); C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) at 101:19–102:16 ("[A]ll the known copyrighted books with all the known copyrighted claims of people were removed from index; not only from the subset that sent to OpenAI, but even from the index that Microsoft had on -- on its -- Bing website."); *see also* ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23. Defendants further dispute any implication that Microsoft knowingly and/or willfully provided Bing Index data to OpenAI that contained copyrighted books. Defendants further dispute any implication from the claim that Microsoft transferred portions of Bing Index to OpenAI only "[a]fter receiving the data it sought from OpenAI" in Paragraph 593 that Microsoft's transfer of portions of the Bing Index was conditioned on its receipt of GPT-3 training data from OpenAI. Books Ex. 1 (K. Allison

	30(b)(1) Tr.) at 72:1–8 [redacted]); *see also id.* at 68:20–23 [redacted]). Further, the facts stated in Paragraph 593 are immaterial because the cited evidence does not support that the Asserted Works were in either the portion of GPT-3 training data that was transferred to Microsoft or the portion of the Bing Index that was transferred to OpenAI. *See id.*; Books Ex. 275 at OPCO_MDL_001653588 (listing only commoncrawl, enwiki, and webtext2 as the training datasets shared in November 2021 and not the LibGen datasets). These facts therefore do not affect the outcome of the suit under the governing law.
594. [redacted]	594. **Disputed in Part – Not Supported by Cited Evidence; Improper Legal Argument.** Microsoft responds as follows: Regarding the first sentence, Microsoft does not dispute that Mr. Ribas acknowledged those statements at his deposition. C. Blythe Opp. Decl. Ex. 84 (J. Ribas 30(b)(1) Tr.) at 179:17–180:21. Microsoft also does not dispute that Microsoft has processes to filter out known copyrighted materials contained in the Bing Index without the copyright owner's authorization. *See* C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 52:15–53:15 ("So we're aware the internet has all kinds of information. That said, we do the best diligence we can to

	go after content which we think is legitimate. But there is no guaranteed way to do it but for following certain guardrails and guidelines and policies thereafter. So while there's a possibility, we under- -- if there is a copyrighted page, we don't even want to show it in search. That's why we take the whole DMCA process very seriously. And we have a team dedicated. We have a set of people who are, at a very time sensitive manner, removing such content from the index. So Microsoft's stance is that we have put enough diligence, both during the data collection process and on a continuous basis, to make sure that the index doesn't contain copyrighted content to the best possible extent."); ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23. But Microsoft disputes any implication in Paragraph 594 that it did not take measures to prevent or remove known unauthorized copyrighted content from the Bing Index data it provided to OpenAI. C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) at 101:19–102:16 ("[A]ll the known copyrighted books with all the known copyrighted claims of people were removed from index; not only from the subset that sent to OpenAI, but even from the index that Microsoft had on -- on its – Bing website."); ECF 1705 (P. Gaggara Decl.) ¶¶ 20–22. Regarding the second sentence, Microsoft further disputes this statement to the extent that it implies any such "legal risks" relate to copyright. *See* C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr. (Feb. 10, 2026)) at 55:11–21

	("[D]ata is a very, very broad concept, right, that is -- encompasses numerous things that intersects with a whole variety of sort of various laws and regulations. Personally identifiable data in the intersection of privacy. Privacy is obviously a, you know, evolving area of the law. Data sovereignty and data residency and where data resides and its ability to move across geographic borders is obviously a rapidly evolving state of the law."). Microsoft also responds that the Court should disregard the second sentence of this Paragraph to the extent that it implies that providing portions of the Bing Index to OpenAI is not legal. That conclusion constitutes improper attorney argument and/or is conclusory. OpenAI defers to Microsoft's response.
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.	595. **Disputed in Part – Not Supported by Admissible Evidence.** Microsoft responds as follows: While Microsoft does not dispute Mr. Ribas made those statements at his deposition, the evidence cited lacks foundation as Mr. Ribas testified that he was not involved with negotiating and had never seen the [redacted] governing OpenAI's use of Bing Index data. *See* C. Blythe Opp. Decl. Ex. 84 (J. Ribas 30(b)(1) Tr.) at 199:5–10. When asked about the purported "restrictions" on OpenAI's use of the Bing Index data, Mr. Ribas testified, "I did not talk to them about this. I was not part of it." *Id.* at 199:18-23. Thus, Mr. Ribas's statements are not

	supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Microsoft further disputes Paragraph 595 to the extent it implies that OpenAI was not subject to restrictions on how it could use the Bing Index data. Microsoft made the Bing Index data available to OpenAI "solely for use by OpenAI and LP for 'computational analysis' ([redacted])." Books Ex. 161 at MSFT_AI_MDL_000097041–42; *see also* C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 93:5–95:2 (explaining that the term "computational analysis" means [redacted]). OpenAI defers to Microsoft's response.
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.	596. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that OpenAI deleted the Bing Index. *See* C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 80:3–6; ECF 1814–1 (S. Shan Decl. Ex. A) ¶ 128; C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 346:6–13; Books Ex. 170 at MSFT_AI_MDL_000235059. Defendants dispute any implication that OpenAI trained its large language models using the Bing Index data that it received from Project Taxi because it is not supported by the cited

	evidence. Specifically, Qiming Yuan is not an OpenAI employee "who trained models on the Bing Index data." Mr. Yuan testified that he was "not aware of OpenAI using Taxi data to train any model" and that he "didn't use Taxi data [Bing Index data] to train GPT models" because he "didn't get the permission to do so" from his manager. *See* C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 80:7–20.

B. Microsoft Transferred Crawled Data to OpenAI (Project Mango)

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.	597. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: While the cited evidence does not state that "Microsoft declined to permit OpenAI to use the Bing Index data to train its publicly available models," Defendants do not dispute this statement. Under the terms of the 2019 Statement of Work governing OpenAI's use of Bing Index data, Microsoft made the Bing Index data available to OpenAI solely for "computational analysis," meaning that OpenAI was only to use the data [redacted]; the scope of permitted uses did not change at any later time. Books Ex. 161 at MSFT _AI_MDL_000097041–42; C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 143:10–11; *see also id.* at 93:10–95:2. OpenAI disputes that Ms. Gaggara, a Microsoft employee, is authorized to speak on its behalf about OpenAI's decision-making. Defendants also do not dispute that OpenAI decided to conduct its own crawl of the internet but deny any causal

	connection between Microsoft's decision to not permit OpenAI to use the Bing Index for training publicly available models and OpenAI's decision to conduct its own crawl, and the cited evidence does not support any such causal connection.
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.	598. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Defendants dispute that Microsoft transferred "a list of URLs" to OpenAI because that statement is not supported by the cited evidence. Microsoft provided OpenAI a list of domains (*i.e.*, the name of the website such as www.example.com), not URLs (*i.e.*, addresses for specific webpages within a website). *See* C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 112:19–24. Microsoft did not have visibility into the specific content OpenAI obtained from the domains from which Microsoft crawled content on OpenAI's behalf or from the domains from which OpenAI itself directly crawled content. *See* C. Blythe Opp. Decl. Ex. 65 (Gaggara Jan. 27, 2026 30(b)(6) Depo. Tr.) at 36:6-14. Microsoft also disputes the assertion that Microsoft provided the list of domains "to guide OpenAI's crawler"; Ms. Gaggara, Microsoft's 30(b)(6) witness on topics concerning the web crawling services Microsoft provided to OpenAI, testified: "It was always supposed to be OpenAI which was guiding, directing what had to happen with that crawler." C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 155:21–23.

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.	599. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Defendants do not dispute that OpenAI lacked crawling infrastructure and search-engine experience at some point in time, or that Microsoft had crawling infrastructure and search-engine experience. But Defendants dispute that Microsoft supplied "seeding URLs" to OpenAI. Microsoft provided OpenAI a list of domains, not URLs. *See* C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 112:19–24. Defendants further dispute any implication from the claim that Microsoft provided OpenAI a list of "URLs" that Microsoft provided OpenAI a list of specific works associated with those URLs to crawl; the list of domains that Microsoft provided did not point to any specific content or works that may (or may not) have been contained in the webpages within those domains, and Microsoft did not have visibility into the specific works OpenAI obtained from the domains from which Microsoft crawled content on OpenAI's behalf or from the domains from which OpenAI itself directly crawled content. *See* C. Blythe Opp. Decl. Ex. 65 (Gaggara Jan. 27, 2026 30(b)(6) Depo. Tr.) at 36:6-14. Defendants also do not dispute that Microsoft's guidance regarding crawling was helpful to OpenAI. But Defendants dispute that Microsoft's provision of a list of domains was "necessary for OpenAI's crawl to be effective" because that assertion is not supported by the cited evidence. None of the cited evidence states that the crawling services were "necessary." Rather, the cited evidence

	describes certain guidance Microsoft provided, but it does not state that OpenAI could not have conducted a crawl without Microsoft's guidance. *See* C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 214:19–215:14; C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 184:237–185:10; Books Ex. 170 at MSFT_AI_MDL_000235059. Additionally, Ms. Gaggara testified, "Microsoft offered technology. But if not Microsoft, they could write the code on their own. It might take a few more months." C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 148:23–149:17.
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.	600. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that they entered a contract in October 2023 for web crawling services. Books Ex. 783 at OPCO_MDL_004771618. Defendants dispute that "OpenAI would crawl half of the web" and "Microsoft would crawl the other half" because Exhibits 783 and 57(a) do not support this assertion. Microsoft (on behalf of OpenAI) and OpenAI itself each crawled URLs from half of a subset of domains, not all domains that exist on the web. Books Ex. 783 at OPCO_MDL_004771618; C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 143:25–144:1 ("We would run a service on their behalf to crawl the URLs…."); C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 34:19–38:13; *see also* C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 108:15 ("It's tiny versus the whole internet."). Microsoft did not crawl for OpenAI any

	webpages that used robots.txt directives indicating that the web publisher did not want their content crawled by GPTBot. C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 144:3–5; C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 96:21–97:10. Additionally, Books Ex. 783 makes clear that the web crawling services Microsoft provided were subject to various restrictions and limitations. Books Ex. 783 at OPCO_MDL_004771618; C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 336:20–344:1. Defendants also dispute any implication in Paragraph 600 that Microsoft profited from this agreement. Any payments made by OpenAI to Microsoft in connection with its crawling services were purely for cost recovery. *See* C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 198:19–199:2 ("[I]t is not revenue in layman terms because we are not making a profit here. It is pure cost."); C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 146:3–7; C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 343:24–344:1 ("the agreement is more about the operational burden of running a web-crawling service rather than paying for content."). Microsoft further disputes the statement that Microsoft "transfer[red] the downloaded content" to OpenAI. The crawled content was streamed directly to OpenAI's Azure storage container, which Microsoft could not access. C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20,

	2025)) at 145:4–9; C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 31:17–19.
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.	601. **Undisputed.**
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.	602. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 602 are disputed because they are not supported by the cited evidence. Specifically, the cited evidence does not state that OpenAI "decided to crawl the full internet," nor does it describe any resulting data as "'GPT-bot' data." *See* C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 456:8–16. Microsoft defers to OpenAI.
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.	603. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants respond as follows: Defendants do not dispute that Dr. Shan testified to the statements in Paragraph 603. But Defendants respond that Paragraph 603 is disputed on the grounds that it is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the evidence cited as Books Exhibit 53 and quoted in Paragraph 603 that "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'" lacks foundation. C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) at 309:2–24. OpenAI defers to Microsoft's response.
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.	604. **Disputed in Part.** Microsoft responds as follows: Regarding the first sentence, Microsoft does not dispute that it began efforts to initiate the Project Mango crawling services around August 2023 and began streaming the crawled data to OpenAI around September 2023. But Microsoft responds that the crawled content was streamed directly to OpenAI's Azure storage container, not otherwise "transferred." C. Blythe Opp. Decl. Ex. 65 P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026) at 31:17–19. Regarding the second sentence, Microsoft does not dispute that the domains libgen.ee, thepiratebay.ee, zlibrary.cc, and e-books.com were included in the list of the domains to be crawled by Microsoft. But Microsoft does dispute any implication in Paragraph 604 that Microsoft crawled and collected content from the domains libgen.ee, thepiratebay.ee, zlibrary.cc, and e-books.com. Microsoft used the existing codebase for Bingbot, *see* C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 141:4–8, which is a compliant crawler that does not retrieve content that is only available through torrent protocols. ECF 1705 (P. Gaggara Decl.) ¶¶ 20–22. Additionally, the existence of a URL on the web does not mean that it is

	accessible to the crawler and/or downloadable content exists on the URL's associated webpage. *See* C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 66:5–69:5. Further, Plaintiff expert Shan did not identify any content from these domains in OpenAI's GPTBot dataset purportedly crawled by Microsoft. *See* ECF 1814-1 (S. Shan Decl. Ex. A) ¶ 137; (S. Shan Decl. Ex. A) Appendix L. OpenAI defers to Microsoft's response.
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.	605. **Disputed in Part – Incomplete.** Microsoft does not dispute that Mr. Canel made those statements at his deposition. But, pursuant to Fed. R. Evid. 106, Mr. Canel later testified: "Let's say some sites provide data to Microsoft and they may not want to provide data for partners of Microsoft. So as we said, enabling this crawl allowed to control access or no access to the content" and "you ask if we configure the software name Mango to take into account the GPTBot robots block . . . And the answer is yes, we were crawling the web on behalf of OpenAI following the robots protocol and blocking GPTBot if webmasters were – didn't want to crawl the web for GPTBot." C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 78:5–9; *id.* at 96:21–97:10. Microsoft further disputes any implication in Paragraph 605 that Microsoft did not follow responsible crawling practices in connection with the crawling services it provided to OpenAI. Microsoft used the existing codebase for Bingbot, *see* C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 141:4–8, which is a

	compliant crawler that does not retrieve content that is only available through torrent protocols. ECF 1705 (P. Gaggara Decl.) ¶¶ 20–22. Microsoft further responds that Microsoft dedicates substantial efforts to prevent the inclusion of and diligently remove any copyrighted content contained in the Bing Index without the copyright owner's permission, and likewise to ensure that such content was not provided to OpenAI. C. Blythe Opp. Decl. Ex. 65 P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026) at 52:15–53:15 ("So we're aware the internet has all kinds of information. That said, we do the best diligence we can to go after content which we think is legitimate. But there is no guaranteed way to do it but for following certain guardrails and guidelines and policies thereafter. So while there's a possibility, we under- -- if there is a copyrighted page, we don't even want to show it in search. That's why we take the whole DMCA process very seriously. And we have a team dedicated. We have a set of people who are, at a very time sensitive manner, removing such content from the index. So Microsoft's stance is that we have put enough diligence, both during the data collection process and on a continuous basis, to make sure that the index doesn't contain copyrighted content to the best possible extent."); ECF 1705 (P. Gaggara Decl.) ¶¶ 20–23. OpenAI defers to Microsoft's response.

606. Microsoft intended to keep Project Mango "secret." 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.	606. **Disputed – Incomplete.** Microsoft responds as follows: Regarding the first sentence, Microsoft disputes any implication that it "intended to keep Project Mango 'secret'" for improper reasons. First, Microsoft was acting in its capacity as a technology service provider and did not disclose the activities of its customers consistent with typical business practices. C. Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Jan. 27, 2026)) at 142:10–145:11 ("There was no attempt to keep it [Project Mango] secret or make it public. It was a technology we were offering OpenAI, and OpenAI was using it."). Second, to the extent websites had a complaint or questions regarding crawling activities, those questions would be properly directed toward OpenAI as the party responsible for the crawling activity. C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 85:25–86:19. Microsoft also disputes the characterization that "[a]t least one Microsoft employee objected to Microsoft's role in Project Mango" for any reason related to the issues in the case. Pursuant to Fed. R. Evid. 106, Mr. Canel testified that he "wanted to revisit the complexity of crawling" because "I'm very busy at Microsoft. This is an additional project for me. That's why every time there is a new project to make sure that we do appropriate prioritization of work." C. Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) 77:18–25. OpenAI defers to Microsoft's response.

C. OpenAI Transferred LibGen to Book-Summarization Contractors

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.	607. **Disputed – Not Supported by Cited Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 607 are disputed because they are not supported by the cited evidence. Specifically, OpenAI did not distribute LibGen content to third-party contractors. The contractors were permitted only to view excerpts of books one at a time through OpenAI's website interface. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 127:7–23. Further, Wu testified [redacted] *Id.* at 121:23–122:8. Ouyang also testified that the specific techniques discussed in Exhibit 18 were not used to train OpenAI's commercial models. C. Blythe Opp. Decl. Ex. 59 (L. Ouyang 30(b)(1) Tr.) at 119:7–20. [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 21:4–16. Microsoft defers to OpenAI.
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.	608. **Disputed in Part**. OpenAI responds as follows: Undisputed that text materials were used as part of the book summarization project. However, disputed that Plaintiffs' broader characterization and inferences drawn

	from Wu's testimony—[redacted]—are corroborated by the cited evidence. Wu testified [redacted]. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 110:12–111:4. Microsoft defers to OpenAI.
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.	609. **Disputed**. OpenAI responds as follows: Paragraph 609 is disputed because Plaintiffs' characterization of and inferences drawn from Wu's testimony are inaccurate and incomplete. Pursuant to Fed. R. Evid. 106, OpenAI responds that Wu testified that [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 87:7–88:7. The cited testimony also does not support the assertion that Wu was "the researcher primarily responsible for the LibGen portions of the project." *Id.* at 86:4–87:6. Microsoft defers to OpenAI.
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.	610. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that Wu retained LibGen-derived training datasets. However, disputed that the materials Wu retained

	constituted "books." *See* B. Frederiksen-Cross Opp. Decl. Ex. A ¶ 149 ("Even a cursory examination of these files reveals a notable amount of non-book content—e.g., poems, catalogs, and pamphlets, author biographies, and academic papers."). Further, disputed that "OpenAI deleted the full set of files used in this project." The cited portion of Shan's expert report, S. Shan Decl. Ex. A ¶ 72, does not support the proposition that OpenAI deleted the full set of files used in this project. The disputed portions of Paragraph 610 are therefore unsupported by the record. Microsoft defers to OpenAI.
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.	611. **Disputed – Not Supported by Cited Evidence**. OpenAI responds as follows: The facts set forth in Paragraph 611 are disputed because they are not supported by the record. Specifically, the cited testimony does not support Plaintiffs' assertion. Ouyang testified that he "do[esn't] know" if there were even records created "that would show who accessed the server and when" at the time the books were available to contractors. C. Blythe Opp. Decl. Ex. 59 (L. Ouyang 30(b)(1) Tr.) at 185:14–186:15. Further, OpenAI did not distribute LibGen content to third-party contractors that could have been "retained." [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 127:7–23. Microsoft defers to OpenAI.

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.	612. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: Undisputed that by 2018, Google had emerged as a company developing artificial intelligence. However, disputed that Plaintiffs may rely on the cited third-party article to establish that Google was the "leading company in artificial intelligence" by 2018. The cited article is an out-of-court statement offered for the truth of that assertion and constitutes inadmissible hearsay. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Further disputed that the article's characterization of Google as the "leading" AI company states an objective fact rather than the author's opinion or characterization. For those reasons, the Court should disregard Paragraph 612 because it is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. However, the fact stated in Paragraph 612 is immaterial because this fact does not affect the outcome of the suit under the governing law. Specifically, Google's standing in artificial intelligence in 2018 has no bearing on the claims or defenses asserted in this case.
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	613. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 613 are disputed because they are

Decl. Ex. 2(a) at 25:7–26:2; Nelson Decl. Ex. 32 at 91:21–92:9.	unsupported by the record. Disputed that Plaintiffs' characterization that OpenAI employed "few people" and had "little money to buy the data and compute its technology needed" is supported by the cited evidence. Books Ex. 2(a) (S. Altman 30(b)(1) Tr.) at 25:14–18 states that OpenAI had approximately 40 researchers and "very little resources," and C. Blythe Opp. Decl. Ex. 86 (U. Madan 30(b)(1) Tr.) at 91:25 states that OpenAI "was at the cutting edge of AI research," but neither states the fact proposition for which the testimony is cited. Microsoft defers to OpenAI.
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.	614. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Microsoft responds as follows: Microsoft does not dispute that, in 2019, Kevin Scott and Jon Tinter viewed Google and OpenAI as ahead of Microsoft in large-scale AI model training. Microsoft disputes that it "held more resources and compute than most companies" because the cited evidence does not support that assertion. To the contrary, the cited evidence reflects that Microsoft lacked AI training infrastructure at the scale of Google's. [redacted]

	[redacted]. *See* Books Ex. 157 at MSFT_AI_MDL_000011364; Books Ex. 160 at MSFT_AI_MDL_000050904; C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 367:16–368:3. Microsoft also responds that the facts in Paragraph 614 are immaterial because Google's standing in artificial intelligence has no bearing on the claims or defenses asserted in this case. OpenAI defers to Microsoft's response.
615. In 2016, Google publicly declared itself an "AI First" company. Nelson Decl. Ex. 75 at 4.	615. **Disputed – Not Supported by Admissible Evidence; Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 615 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the cited Founder's Letter at Books Ex. 75 is an out-of-court statement offered for the truth of the assertion and constitutes inadmissible hearsay. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). The facts set forth in Paragraph 615 are further disputed because they are not supported by the cited evidence. Finally, disputed that Books Ex. 75 supports Plaintiffs' characterization that Google publicly declared itself an "AI First" company in

	2016. It discusses the evolution of computing and states prospectively that "[w]e will move from mobile first to an AI first world." It does not state that Google was or was becoming an "AI First" company. Further, the fact stated in Paragraph 615 is immaterial because this fact does not affect the outcome of the suit under the governing law. Specifically, Google's standing in artificial intelligence has no bearing on the claims or defenses asserted in this case.
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.	616. **Undisputed.** Microsoft defers to OpenAI.
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.	617. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the fact stated in Paragraph 617 is immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.
618. 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	618. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that Paragraph 618 is immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue.

at 1. [redacted]	OpenAI defers to Microsoft's response.
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.	619. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the fact stated in Paragraph 619 is immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.
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. [redacted]	620. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the fact stated in Paragraph 620 is immaterial because Google's technology, assets, and position are not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.
621. [redacted]	621. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the facts related to Google stated in Paragraph 621 are immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue.

[redacted]	OpenAI defers to Microsoft's response.
622. [redacted]	622. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the facts related to Google stated in Paragraph 622 are immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.
623. [redacted]	623. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the facts related to Google in Paragraph 623 are immaterial because Google's standing in artificial intelligence is not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.

[redacted]	
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.	624. **Undisputed – Incomplete; Immaterial.** Microsoft does not dispute that Books Exhibit 40 contains the quoted language, and that OpenAI used Google's cloud platform for computing before entering into the Microsoft partnership. But pursuant to Fed. R. Evid. 106, in the same answer cited, Mr. Nadella testified that, when OpenAI "w[as] just a nonprofit, they were looking for free compute. We were their first compute provider. Then they moved, I think, to Amazon, then they moved to Google." Books Ex. 40 (S. Nadella 30(b)(1) Tr.) 248:13–25. Microsoft also responds that the facts related to Google in Paragraph 624 are immaterial because OpenAI's prior use of Google's cloud services is not relevant to the claims and defenses at issue. OAI defers to Microsoft's response above, and also responds as follows: Disputed to the extent Plaintiffs seek to use Nadella's testimony for the truth of the matter about OpenAI's "realization[s]." The disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, Nadella is a Microsoft employee not authorized to speak on behalf of OpenAI. The evidence cited constitutes inadmissible hearsay, and Nadella's testimony is

	objectionable for lack of foundation and personal knowledge as to OpenAI's internal deliberations.
625. An OpenAI presentation on the Microsoft partnership listed as its first goal: [redacted] Nelson Decl. Ex. 790 at Slide 2. The presentation stated that [redacted] *Id.* at Slide 4.	625. **Undisputed – Immaterial.** This fact is undisputed but Microsoft responds that the fact related to Google in Paragraph 625 is immaterial because OpenAI's description of itself as a "counterweight" to Google and DeepMind is not relevant to the claims and defenses at issue.
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.*	626. **Undisputed.** Microsoft defers to OpenAI.
627. Microsoft acquired an ownership stake in OpenAI, invested billions of dollars, and placed engineers at OpenAI. *See infra* §§ VI.B, VI.E. OpenAI gave Microsoft access to its technology and training data. *See infra* §§ V.A, VI.E.	627. **Disputed in Part – Not Supported by Cited Evidence.** Microsoft responds as follows: Microsoft does not dispute that it has invested in, and has a contractual relationship with, OpenAI. But the cited support references "*infra* §§ VI.B, VI.E" which does not itself constitute evidence. Thus, the facts set forth in Paragraph 627 are disputed because they are not supported by the cited evidence. To the extent the facts asserted in

	those sections are individually disputed, so too is this summary assertion. Further, Microsoft disputes the characterization in Paragraph 627 that Microsoft "placed engineers at OpenAI." Under the JDCAs, a limited number of Microsoft engineers worked onsite at OpenAI, took "business direction from OpenAI," and served as "a primary means to accomplish Commercialization Support." Books Ex. 146(b) at MSFT AI MDL 000000599. [redacted] C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 239:24–240:19; C. Blythe Opp. Decl. Ex. 54 (Templeton 30(b)(6) Tr.) at 100:4-12. Microsoft further disputes the assertion that OpenAI "gave Microsoft access to its … training data" as unsupported by the cited evidence. The JDCA gave [redacted] C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 236:17–238:16; *see also* Books Ex. 40 (S. Nadella 30(b)(1) Tr.) at 61:5–62:4 ("Even though we have IP, we don't have access to all the training data" because OpenAI's training process "is always

	opaque" and [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.	628. **Disputed in Part – Not Supported by Admissible Evidence; Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: Defendants do not dispute that Google built a digital database of scanned books in connection with the Google Books project. But Defendants respond that the statement "Google's lead extended to books" or that Google "spent more than a decade building a digital database of approximately 40 million scanned books" is disputed because the cited evidence does not support the claims. The cited evidence does not support the assertion that Google held any "lead" over anyone. Books Ex. 76 at 1. Defendants also respond that the facts set forth in Paragraph 628 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. S.D.N.Y. Specifically, Books Ex. 76, is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. And *Authors Guild v. Google, Inc.*, 804 F.3d 202, 208 (2d Cir. 2015) is case law, not evidence.

	Defendants also respond that the facts set forth in Paragraph 628 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, the size or history of a non-party's book-digitization project has no bearing on the claims or defenses asserted in this case.
629. [redacted]	629. **Undisputed.**

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

630. In July 2018, Microsoft first considered investing in OpenAI to provide the compute OpenAI needed for AI development. Nelson Decl. Ex. 791 at -763–64. Microsoft declined because it was uncertain how closely OpenAI's goals aligned with Microsoft's "business interests." *Id.* at -763.	630. **Disputed in Part – Not Supported by Cited Evidence.** Microsoft does not dispute that it was evaluating a potential investment in OpenAI in mid-2018 or that Ex. 791 contains the quoted language. But Microsoft disputes that it considered investing "to provide the compute OpenAI needed," because the cited evidence does not support those assertions. Microsoft disputes that the exhibit reflects a proposal that Microsoft "declined," or a determination that OpenAI's goals did not align with Microsoft's business interests generally. The exhibit is an email paraphrasing [redacted] Books Ex. 791 at MSFT_AI_MDL_000011763. The quoted language addresses only [redacted]

	[redacted] *Id.* OpenAI defers to Microsoft's response.
631. In February 2019, Microsoft CTO Kevin Scott wrote to CEO Satya Nadella expressing concern about Microsoft's AI training infrastructure. Nelson Decl. Ex. 792 at -361–63. [redacted] *Id.* at -361. [redacted] *Id.*	631. **Undisputed – Immaterial.** Microsoft responds that the facts related in Paragraph 631 are immaterial because Microsoft's 2019 internal assessment of its own AI training infrastructure does not affect the outcome of the suit under the governing law. Microsoft further responds that the facts related to Google in Paragraph 631 are immaterial because Google's activities and capabilities are not relevant to the claims or defenses at issue. OpenAI defers to Microsoft's response.
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.	632. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that in February 2019, OpenAI published "Language Models are Unsupervised Multitask Learners," describing GPT-2. But Defendants dispute the statement in Paragraph 632 that the paper's publication alone "led to" that meeting. Plaintiffs' characterization is inaccurate and misleading because it implies a direct and singular causal link between the paper's publication and these subsequent meetings that is not supported by the cited evidence. Before the paper was published, the meeting was already scheduled for OpenAI to show Mr. Nadella its work. *See* Books Ex. 792 at MSFT _AI_MDL_000010361–62; Books Ex. 182 at

	MSFT_AI_MDL_000892073, MSFT_AI_MDL_000892085. Defendants also respond that the statement in Paragraph 632 that OpenAI "disclosed its use of LibGen" at the demonstration is disputed because it is not supported by the cited evidence. Neither Exhibit 789 nor Exhibit 182 supports the claim that OpenAI "disclosed its use of LibGen" at the demonstration. *See* Books Exs. 182 at MSFT_AI_MDL_000892072, 789. Additionally, Mr. Gates testified that he has "never heard of Lib Gen." C. Blythe Opp. Decl. Ex. 14 (W. Gates 30(b)(1) Tr.) at 115:6-116:14, 119:2-14, 123:12-124:9. Microsoft further responds that the facts in Paragraph 632 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion.
633. By April 24, 2019, Microsoft had decided to proceed with the partnership. Nelson Decl. Ex. 206. Kevin Scott wrote: [redacted] *Id.*	633. **Undisputed.**
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.	634. **Disputed in Part.** Microsoft does not dispute that it agreed to invest $1 billion in OpenAI in July 2019, made an incremental $2 billion commitment in 2021, and committed an additional $10 billion in 2023, for a total

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.	commitment of $13 billion. But Microsoft disputes any implication in Paragraph 634 that those commitments were cash investments. The 2019 and 2021 commitments required minimum Azure spend of [redacted] *See* Books Ex. 147 at MSFT_AI_MDL_000000735. Of the $13 billion total, [redacted] *See* Books Ex. 203 at MSFT_AI_MDL_001550855. OpenAI defers to Microsoft's response.
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).	635. **Undisputed.**
636. [redacted]	636. **Undisputed.**
637. On July 2, 2019, Microsoft and OpenAI entered into a Joint Development and Collaboration	637. **Undisputed.**

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.	638. **Undisputed.**
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.	639. **Undisputed**.
640. The JDCA included collaboration on designing and building "(i) a supercomputer processing unit to be operational [redacted] and "(ii) a supercomputer expected to be operational later in the Term," defined as the "AGI Relevant Supercomputer." Nelson Decl. Ex. 145 at -353.	640. **Undisputed.**
641. [redacted]	641. **Undisputed.**

642. The JDCA prohibited the parties from exchanging data that "(i) contains any personally identifying information, without first obtaining the receiving party's advance written consent; (ii) the disclosing party does not have all rights and consents necessary to disclose such data to the receiving party; or (iii) does not comply with all applicable Law." Nelson Decl. Ex. 145 at -354.	642. **Undisputed.**
643. The same prohibition appeared in each subsequent version of the agreement. Nelson Decl. Ex. 150 at -083; Nelson Decl. Ex. 146 at -589.	643. **Undisputed.**
644. [redacted]	644. **Undisputed.**
645. The JDCA established a Governing Board with equal numbers of Microsoft and OpenAI representatives, meeting at least monthly to manage "the parties' relationship under this Agreement," including "providing advice and recommendations with respect to the overall management of, and strategic direction for, their performance." Nelson Decl. Ex. 145 at -361, -372-373.	645. **Undisputed.**
646. In March 2021, the parties signed the Amended and Restated JDCA. [redacted]	646. **Undisputed**.

[redacted] Nelson Decl. Ex. 150 at -084-085.	
647. [redacted]	647. **Undisputed.**
648. [redacted]	648. **Undisputed.**
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.	649. **Disputed in Part.** The cited support references "*infra* § VI.G" which does not itself constitute evidence; to the extent the facts asserted in those sections are individually disputed, so too is this summary assertion.
650. The Second Amended JDCA provided for immediate delivery of all Licensed IP to Microsoft "[a]t or prior to the first meeting of the Governing Board after the Effective Date." "Licensed IP expressly includes all IP and Technology of OpenAI, OpCo, or both, excluding only AGI." Nelson Decl. Ex. 146 at -581, -586.	650. **Disputed in Part.** Microsoft does not dispute that Books Ex. 146 contains the quoted language. But Microsoft disputes that the Second Amended JDCA provides for "immediate delivery of all Licensed IP." Section 3(a) provides that OpenAI "will disclose and describe" Licensed IP through "a high-level catalog of Models and development projects . . . and a brief description of each," and that Licensed IP and Technology "will be Made

	Available only on Microsoft's written request." Books Ex. 146(a) at MSFT_AI_MDL_000000586–87. OpenAI defers to Microsoft's response.
651. The Second Amended JDCA created a Data Working Group, constituted by the Governing Board, to "regularly discuss and regularly manage issues and decisions related to the acquisition and use of data." Nelson Decl. Ex. 146 at -589.	651. **Disputed in Part – Not Supported by Cited Evidence; Incomplete.** Microsoft does not dispute that the Second Amended JDCA contains the quoted language. Microsoft disputes the characterization that the Second Amended JDCA "created [the] Data Working Group" as unsupported by the cited evidence. The cited evidence states: "As soon as practicable after the Effective Date, the Governing Board *will* constitute a permanent working group . . . ." Books Ex. 146(b) at MSFT_AI_MDL_000000589 (emphasis added). The cited evidence does not state that the Data Working Group was ever created or constituted. Further, Microsoft disputes any implication in Paragraph 651 that the Data Working Group was to discuss and manage "issues and decisions related to the acquisition and use of data" writ large. Pursuant to Fed. R. Evid. 106, Microsoft responds that Paragraph 651 truncates the key part of the sentence, which states: "the Governing Board will constitute a permanent working group to regularly discuss and regularly manage issues and decisions related to the acquisition and use of data *under or in relation to this Agreement*." Books Ex. 146(b) at MSFT_AI_MDL_000000589 (emphasis added). The

	Agreement further provides that "OpenAI may train all Models using data selected solely by . . . OpenAI (although . . . OpenAI will, at all times, consider in good faith the recommendations promulgated by the Data Working Group)." *Id.* at MSFT_AI_MDL_000000590. OpenAI defers to Microsoft's response.
652. Under the three JDCAs, Microsoft's compensation took the form of capped redemption amounts totaling $92 billion: $20 billion under the first, $12 billion under the second, and $60 billion under the third. Nelson Decl. Ex. 60 at 325:5–25; Nelson Decl. Ex. 203 at 861.	652. **Disputed in Part – Not Supported by Cited Evidence.** Microsoft does not dispute that Nelson Decl. Ex. 203 lists Target Redemption Amounts of $20 billion, $12 billion, and $60 billion for Microsoft or that Microsoft and OpenAI entered into three JDCAs. Microsoft disputes that those amounts arose "[u]nder the three JDCAs," because the cited evidence does not support that assertion. Ex. 203 lists the Target Redemption Amounts by Microsoft's investment "Round," not by JDCA, and states that each is "determined based on a negotiated multiple of the investment made." Books Ex. 203 at MSFT_AI_MDL_001550862. The Target Redemption Amounts arise under the parties' investment agreements, not the July 2019 JDCA or its amended and restated versions. *See id.*; Books Exs. 145 at MSFT_AI_MDL_000000345, 150 at MSFT_AI_MDL_000004071, 146 at MSFT_AI_MDL_000000575. Further, Nelson Decl. Ex. 203 is a draft Microsoft memo that does not identify the author or recipient ("To:"

	and "From:" are blank), and therefore cannot be substantiated as containing correct information and is thus not admissible evidence, as required by Local Rule 56.1. OpenAI defers to Microsoft's response.

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.	653. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants respond that Paragraph 653 is disputed in part as to the first sentence and incorporate by reference the responses to Paragraphs 187 and 190. Defendants further dispute Paragraph 653 with respect to the assertion that "Microsoft's knowledge extended beyond that initial disclosure." That statement is not a statement of fact but rather Plaintiffs' characterization and argument summarizing the facts that follow. *See* SUF ¶¶ 654–669. Moreover, Paragraph 653 does not cite independent evidentiary support—it cross-references "supra § II.D and infra this section." Microsoft further responds that the facts set forth in Paragraph 653 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' motion for summary judgment.
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	654. **Disputed in Part – Immaterial; Incomplete.** Microsoft does not dispute that Umesh Madan received the document cited as Books Ex. 189 and

"~11B words from Library Genesis" to its training data. Nelson Decl. Ex. 188 at -958; Nelson Decl. Ex. 189 at -964.	that this exhibit contains the quoted language. But Microsoft responds that the facts set forth in Paragraph 654 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' motion for summary judgment. Additionally, pursuant to Fed. R. Evid. 106, Microsoft further responds that Mr. Madan testified that he "did not read this document," and did not know what LibGen was at the time of receiving this document. C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:7–22. OpenAI defers to Microsoft's response.
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.	655. **Disputed in Part – Not Supported by Admissible Evidence.** Microsoft does not dispute that Books Ex. 189 includes a link to a Wikipedia page describing LibGen. Microsoft does dispute that Paragraph 655 is supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, Books Ex. 795 is hearsay and lacks foundation because it reflects anonymously edited Wikipedia revisions that recite third-party reports. OpenAI defers to Microsoft's response.
656. Madan, who attended the April 2019 Gates demonstration, testified that OpenAI confirmed it had	656. **Disputed in Part – Immaterial; Incomplete.** Microsoft does not dispute that Umesh Madan attended OpenAI's April 2019 meeting with Bill

trained its model on books. Nelson Decl. Ex. 661 at 28:12–29:10, 51:23–52:2.	Gates ("the April 2019 Gates demonstration"). But Microsoft responds that the facts set forth in Paragraph 656 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. Additionally, pursuant to Fed. R. Evid. 106, Microsoft responds that: Mr. Madan testified that he "was not familiar with these data sites" and "didn't know what was in them." C. Blythe Opp. Decl. Ex. 86 (U. Madan 30(b)(1) Tr.) at 59:15–60:8. OpenAI defers to Microsoft's response.
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.	657. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Mr. Madan agreed to the first quoted statement in Paragraph 657, or that he made the second quoted statement at his deposition. Microsoft responds that the facts set forth in Paragraph 657 are immaterial because these facts do not affect the outcome of the suit under the governing law. Specifically, individuals' opinions concerning "piracy" generally are not facts and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.

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.*	658. **Disputed in Part – Incomplete.** Microsoft does not dispute that Exhibit 191 reflects an email from Eric Chung to Umesh Madan that attaches Exhibit 192. But, pursuant to Fed. R. Evid. 106, Microsoft responds that Mr. Madan did not provide deposition testimony concerning Books Exs. 191 and 192, or any versions thereof, but did testify that the first time he had heard of LibGen was during deposition preparation. C. Blythe Opp. Decl. Ex. 15 (U. Madan 30(b)(6) Tr.) at 46:7–16. Eric Chung also testified that he was "unfamiliar with Library Genesis" when he received the paper and as of the date of his deposition, and that he "may not have read the paper in depth to notice that detail." C. Blythe Opp. Decl. Ex. 36 (E. Chung 30(b)(1) Tr. (Feb. 26, 2026)) at 55:12–24. OpenAI defers to Microsoft's response.
659. In February 2020, ahead of a meeting between Dario Amodei and senior Microsoft personnel including Saurabh Tiwary, Mikhail Parakhin, and Dee Templeton, OpenAI sent Microsoft a document titled "Details, Results, and Benchmarking of OpenAI Large Models for MSFT." Nelson Decl. Ex. 696 at -654. The document explained that to address [redacted] OpenAI had [redacted] including 18 billion tokens from "a dataset from Library Genesis (Libgen), which contains books and novels," identified as a [redacted] dataset representing [redacted]. Nelson Decl. Ex. 696.	659. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that OpenAI sent certain Microsoft employees the document reflected in Exhibit 696. But, pursuant to Fed. R. Evid. 106, none of the testimony by Microsoft personnel who testified about receiving the document cited as Ex. 696 demonstrates that Microsoft witnesses had personal knowledge of the sources of data in or contents of LibGen. *See* C. Blythe Opp. Decl. Ex. 51 (S. Tiwary 30(b)(1) Tr.) at 23:3–24:23, 29:18–30:24, 36:6–21; C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) at 54:22–55:5, 61:13–62:6; C. Blythe Opp. Decl. Ex. 54 (D. Templeton 30(b)(6) Tr.) at 285:13–286:23, 291:12–22,

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.	298:9–12**;** C. Blythe Opp. Decl. Ex. 86 (U. Madan 30(b)(1) Tr.) at 54:1–22; C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr.) at 346:7-13. Microsoft further responds that the facts set forth in Paragraph 659 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. OpenAI defers to Microsoft's response.
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.	660. **Disputed in Part – Immaterial; Incomplete.** Microsoft responds that the facts set forth in Paragraph 660 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' summary judgment motion. Additionally, none of the testimony from Microsoft's witnesses about receiving the document cited as Books Ex. 696 demonstrates that Microsoft witnesses had personal knowledge of the sources of data in or contents of LibGen. *See* C. Blythe Opp. Decl. Ex. 51 (S. Tiwary 30(b)(1) Tr.) at 23:3–24:23, 29:18–30:24, 36:6–21; C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr.) at 54:22–55:5, 61:13–62:6; Ex. 54 (D. Templeton 30(b)(6) Tr.) at 285:13–286:23, 291:12–22, 298:9–12**;** C. Blythe Opp. Decl. Ex. 86 (U. Madan 30(b)(1) Tr. at 54:1–22); C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr.) at 346:7–13.

	Defendants dispute Paragraph 660 to the extent this statement characterizes Ms. Murati as having independently acknowledged that the disclosure was made "clearly." In fact, pursuant to Fed. R. Evid. 106, Ms. Murati's full response was, [redacted] *See* Books Ex. 39 (M. Murati 30(b)(1) Tr.) 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.	661. **Disputed in Part – Incomplete.** Microsoft does not dispute that Ms. Templeton agreed to the quoted statement during her deposition. But, pursuant to Fed. R. Evid. 106, Ms. Templeton testified that she was not aware of Library Genesis and had never heard of it before her deposition. C. Blythe Opp. Decl. Ex. 54 (D. Templeton 30(b)(6) Tr.) at 285:13–286:23 (Q: Okay. Microsoft is aware of a shadow library called Library Genesis or LibGen, correct?...The Witness: I'm not aware of it. Q. You're not aware of Library Genesis? A. That's right. Q. Have you ever heard of it before? A. No, not before this."), 291:12–22 ("Q: And what, if anything, did Microsoft do to ensure that there were not copyrighted materials in that training dataset?" "A: We didn't have access to the training dataset."); *id.* at 298:9–12. OpenAI defers to Microsoft's response.
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	662. **Disputed in Part – Incomplete.** Microsoft does not dispute that Ms. Gaggara testified that Microsoft has "no way to know what is copyrighted versus not" in the Bing Index or that Microsoft uses mechanisms to

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]	prevent and/or remove unauthorized copyrighted content in the Bing Index, including DMCA takedown protocols, compliance with robots exclusion directives, and disabling its Bingbot web crawler from retrieving torrented content. C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 150:25–151:5; *see also* C. Blythe Opp. Decl. Ex. 53 (M. Parakhin 30(b)(1) Tr. (Nov. 18, 2025)) at 102:25–103:20 ("The mechanism for Microsoft to know that something is copyrighted is the claim by the owner of the copyright.").
	Microsoft does not dispute that Ex. 144 contains the quoted language in the third sentence. But Microsoft disputes any implication that full or partial copies of books, including the Asserted Works, "were available from Library Genesis." Pursuant to Fed. R. Evid. 106, Sabharwal states in the same email, [redacted] Books Ex. 144 at MSFT_AICPY_000591636–637. The mere existence of a LibGen URL in the Bing Index does not mean that the full content reflected in that URL is stored in the Bing Index. ECF 1705 (P. Gaggara Decl.) ¶¶ 20-22; C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 28:15–31:8.
	Microsoft does not dispute that Ms. Gaggara testified that she was "just learning about LibGen in the context of

	this deposition" despite the fact that she was involved in an email thread from October 2023 that included a reference to LibGen. C. Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 151:7–8. But, pursuant to Fed. R. Evid. 106, when questioned about the email, Ms. Gaggara testified that "she did not even see" or "pay any attention to" the sources themselves. *Id.* at 152:13–25. OpenAI defers to Microsoft's response.
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.	663. **Undisputed – Immaterial.** Microsoft does not dispute the facts stated in Paragraph 663. But Microsoft responds that the facts set forth in Paragraph 663 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Microsoft's training practices are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
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.	664. **Undisputed – Immaterial.** Microsoft does not dispute that Dr. Weyl agreed with the characterization in Paragraph 664 at his deposition. But Microsoft responds that the characterizations set forth in Paragraph 664 are immaterial because they do not affect the outcome of the suit under the governing law. A Microsoft employee's personal opinions based on speculation are not fact and are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.

665. [redacted]	665. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Nelson Decl. Exs. 28 and 187 contain the quoted language. But Microsoft does dispute any implication that Jamie Huynh deposition testimony cited as Ex. 28 were statements made on behalf of Microsoft or provided in Mr. Huynh's capacity as "Microsoft's corporate designee." Topics related to Microsoft's "underst[anding] that downloading copyrighted materials from Library Genesis implicated legal risks" were not in the scope of the 30(b)(6) topics for which he was designated. C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6). Tr. (Jan. 27, 2026)) at 111:15. Additionally, pursuant to Fed. R. Evid. 106, Mr. Huynh testified: "I don't know what our [Microsoft's] formal position was on Library Genesis." *Id.* at 111:6–7. Microsoft further responds that the statements in Paragraph 665 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Specifically, the legality of "downloading copyrighted materials from Library Genesis" in general, and particularly any legal implications of such activities under international law, are immaterial to the issues before the Court. OpenAI defers to Microsoft's response.
666. [redacted]	666. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books

[redacted]	Exhibits 155, 177(a), 177(b), and 59(c) contain the language quoted in Paragraph 666. Additionally, pursuant to Fed. R. Evid. 106, Microsoft responds to Paragraph 666 as follows: With respect to the quoted language from Books Exhibit 155, Ms. Templeton testified that the document [redacted] C. Blythe Opp. Decl. Ex. 54 (D. Templeton 30(b)(6) Tr. (Jan. 20, 2026) at 135:21–137:19). When asked [redacted] Ms. Templeton answered, [redacted] *Id.* at 17:22-18:9. Microsoft further responds that the facts set forth in Paragraph 666 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' summary judgment motion. OpenAI defers to Microsoft's response.
667. [redacted]	667. **Undisputed – Immaterial.** Microsoft does not dispute that Nelson Decl. Exhibit 180 contains the quoted language. But Microsoft responds that the facts set forth in Paragraph 667 are immaterial and therefore

[redacted]	disputed because these facts do not affect the outcome of the suit under the governing law. Specifically, this is a Microsoft internal document speculating what datasets OpenAI "may have used" for training. This is not a fact and immaterial to the issues before the Court. Additionally, the facts set forth in Paragraph 667 are immaterial because Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Book Plaintiffs' summary judgment motion. OpenAI objects to Plaintiffs' use of Books Ex. 180 because it constitutes inadmissible hearsay as to OpenAI. Otherwise, OpenAI defers to Microsoft's response.
668. [redacted]	668. **Disputed in Part.** Microsoft does not dispute that Exhibits 184, 207 and 939 contain the quoted language or that LLMs can at times generate text that was part of their training corpus, a known phenomenon called regurgitation. But Microsoft disputes any implication that the models at issue frequently regurgitate content or that such instances of regurgitation constitute copyright infringement. Instances of regurgitation in a properly trained model are extremely rare. *See* ECF 1723 (J. Lafferty Decl.) ¶ 35; ECF 1707 (J. Usdan Decl.) ¶¶ 19-20. OpenAI defers to Microsoft's response.
669. Sarah Bird, Microsoft's corporate designee, testified that "for a technology like this, I think it's very	669. **Undisputed.**

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

670. Under the Second Amended JDCA, Microsoft agreed to provide OpenAI with 100 percent use of the "Owl" and "Panda" supercomputing processing units, 100 percent use of the Phase 2 Supercomputer, and up to $9 billion for the Argos II Dedicated System and Phase 3 and Phase 4 Supercomputers. Nelson Decl. Ex. 146 at -579, -592–93.	670. **Undisputed.** Defendants respond that Paragraph 670 is undisputed insofar as the contract speaks for itself.
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.	671. **Undisputed.**
672. [redacted]	672. **Undisputed – Incomplete.** Microsoft does not dispute that Nelson Decl. Ex. 36 contains the quoted language. But pursuant to Fed. R. Evid. 106, Mr. Melton testified that the data centers "were at this phase our standard data centers. It's the same thing that we use for the rest of Azure and any other customer," and that "it wasn't a special design of our data centers at this point." Books Ex. 36 (L. Melton 30(b)(6) Tr. (Nov. 4, 2025)) at 35:18–36:7. Mr. Melton agreed that the special design

	aspects were limited to meeting "the GPU count and speed requirements of OpenAI." *Id.* at 36:8-15. And he testified that "[t]he design of them is the same design I use for every customer. It's just a bigger version of it." *Id.* at 87:3-15. OpenAI defers to Microsoft's response.
673. Microsoft dedicated Azure datacenters for OpenAI's exclusive use and provided multiple petabytes of Blob Storage for training data. 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.	673. **Disputed in Part – Not Supported by Cited Evidence; Incomplete; Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Exhibit 27 describes Microsoft's Azure storage and Exhibit 36 contains the quoted language. But Microsoft disputes the characterization in Paragraph 673 of Blob Storage as being provided "for training data" because it is not supported by the cited evidence. Pursuant to Fed. R. Evid. 106, Mr. Melton testified that Blob Storage "is not a AI thing. It's a generic storage solution" that "is used for many things including checkpoints, data, OS," and that Microsoft "didn't know what [OpenAI was] going to use it for" or "how much was going to be used for what requirements." Books Ex. 36 (L. Melton 30(b)(6) Tr. (Nov. 4, 2025)) at 57:7–20; 61:15–62:3. As to the second sentence, Microsoft defers to OpenAI. OpenAI responds as follows: Undisputed that OpenAI has utilized Microsoft Azure infrastructure for model training. However, disputed to the extent Plaintiffs rely on testimony from a Microsoft employee to establish facts about OpenAI's model training operations. The

	assertion that "all GPT models from GPT-3 onward have been trained using Microsoft's Azure supercomputing infrastructure" is a characterization by Microsoft's corporate designee, not by OpenAI. Further disputed that the phrase "from GPT-3 onward" is accurate insofar as it purports to encompass models beyond the Accused Models in scope; the assertion exceeds the scope of Plaintiffs' claims and is immaterial. Further, the disputed proposition in Paragraph 673 is immaterial because it does not affect the outcome of the suit under the governing law. Finally, to the extent Plaintiffs' characterizations of this fact implicate Microsoft's provision of infrastructure, OpenAI defers to Microsoft.
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.	674. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants respond as follows: Disputed because this fact relies exclusively on testimony from a Microsoft employee—Mr. Melton—to establish facts about which OpenAI models were trained on which Azure supercomputing phases. Mr. Melton, as a Microsoft employee, lacks personal knowledge and foundation to testify definitively about OpenAI's internal training operations. Indeed, Mr. Melton testified that Microsoft does not know what OpenAI is training while the training takes place and "learn[s] when they do the release," and as to GPT-3.5 testified only that it "would have been Telemachus." Books Ex. 36 (L. Melton 30(b)(6) Tr.) at 40:30–41:2, 83. Because the disputed proposition is not supported by admissible evidence, as required by Local Rule

	56.1, it is therefore not properly before the Court. S.D.N.Y. Local Rule 56.1(d).
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 ([redacted]).	675. **Disputed.** Defendants respond that Plaintiffs' cited evidence does not establish that the identified storage was "critical" to training the at-issue LLMs. Plaintiffs rely on: (1) deposition testimony from a Microsoft employee (Mr. Melton) who lacks personal knowledge of OpenAI's training infrastructure and operations; and (2) a statement in a Microsoft internal document (Books Ex. 165 at MSFT_AI_MDL_000208548) asserting that [redacted] Mr. Melton testified that storage is "critical for any training. Like every training needs . . . some form of storage, right. [redacted] Books Ex. 36 (L. Melton 30(b)(6) Tr.) at 62–63. He further testified that Microsoft "didn't know what [OpenAI was] going to use it for," only that "some of it had to be for data, some of it was going to be check pointing." *Id.* at 61:22–62:3. Further, the statement in Ex. 165 that [redacted] Books Ex. 165 at –MSFT_AI_MDL_000208548, -549. OpenAI further objects to both sources as inadmissible to establish the fact asserted: Mr. Melton's testimony lacks foundation under Fed. R. Evid. 602, and the

	statement in Ex. 165 is inadmissible hearsay under Fed. R. Evid. 802. Fed. R. Civ. P. 56(c)(2). Further disputed to the extent Plaintiffs characterize a Microsoft employee's statement as establishing facts concerning OpenAI's training infrastructure; the cited statement reflects Microsoft's account, not a representation by OpenAI. Microsoft further responds as follows: The cited evidence does not support that "[t]hat storage" was "critical" to training the at-issue LLMs in any sense specific to those models. First, Mr. Melton testified that storage is "critical for any training. Like every training needs . . . some form of storage, right. [redacted] Books Ex. 36 (L. Melton 30(b)(6) Tr.) at 62–63. He further testified that Microsoft "didn't know what [OpenAI was] going to use it for," only that "some of it had to be for data, some of it was going to be check pointing." *Id.* at 61:22–62:3. Second, the statement in Ex. 165 that [redacted] Books Ex. 165 at –MSFT_AI_MDL_000208548, -549.
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	676. **Undisputed – Incomplete; Immaterial.** Microsoft does not dispute that Nelson Decl. Exhibit 827 contains the quoted language. But Pursuant to Fed. R. Evid. 106, Mr. Russinovich testified that he does "not know" whether "any Library Genesis content is currently stored on

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.	Azure systems," and that he does not "know one way or the other" whether OpenAI's models were trained on Library Genesis data. Books Ex. 827 (M. Russinovich Tr.) at 123:24–124:2, 125:5–8. Microsoft further responds that the statements in Paragraph 676 are immaterial because they do not affect the outcome of the suit under the governing law. A witness's speculation about whether a hypothetical scenario would violate Microsoft's terms of service is not a fact and immaterial to the issues before the Court. To the extent Plaintiffs purport to establish Microsoft's knowledge, Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in the case. OpenAI defers to Microsoft's response.
677. Microsoft also provided Azure supercomputing systems for OpenAI's inferencing operations (Project Mumford), enabling OpenAI to make models available to users. Nelson Decl. Ex. 36 at 85:14–87:2.	677. **Disputed in Part.** Defendants respond as follows: Undisputed that OpenAI utilizes Microsoft Azure infrastructure for inferencing operations. However, the characterization that Microsoft's provision of infrastructure "enabl[ed] OpenAI to make models available to users" is disputed as stated to the extent it implies a but-for causal relationship—i.e., that OpenAI could not have made its models available to users without Microsoft's specific systems. This is Plaintiffs' characterization, not a statement supported by the cited evidence.

678. When OpenAI makes ChatGPT available to users, it uses Microsoft's Azure systems. Nelson Decl. Ex. 36 at 85:14–87:2.	678. **Disputed in Part – Not Supported by Admissible Evidence.** Defendants respond as follows: Undisputed that when OpenAI makes ChatGPT available to users, it uses Microsoft's Azure systems. However, disputed to the extent the cited evidence purports to establish facts about OpenAI's product infrastructure through testimony from a Microsoft employee, Mr. Melton, who lacks personal knowledge of and foundation to testify regarding OpenAI's product infrastructure. For that reason, the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court.
679. Microsoft retained the sole discretion and right—with or without cause—to revoke OpenAI's access to Azure, including the GPUs OpenAI used to train and power its LLMs. Nelson Decl. Ex. 36 at 92:10-98:6; Nelson Decl. Ex. 164 at -502 (Statement of Work Agreement titled "Supercomputer Tech Collateral Exchange with OpenAI," dated June 30, 2020: "Microsoft reserves the right to suspend or revoke access to the Supercomputing Tech, at any time with or without cause.").	679. **Disputed in Part – Not Supported by Cited Evidence.** Defendants do not dispute that Mr. Melton testified that Microsoft has a policy "for any customer" under which it may revoke quota, or that Ex. 164 contains the quoted language. Defendants dispute that the cited evidence shows Microsoft retained "the sole discretion and right" to revoke OpenAI's access to "Azure, including the GPUs OpenAI used to train and power its LLMs." See Books Ex. 164 at MSFT_AI_MDL_000169500-502. The quoted provision applies only to the "Supercomputing Tech" shared under Statement of Work #2, which lists Athena and Apollo silicon specifications and related materials; Paragraph 679 expands the scope of the contractual provision beyond its terms. *Id.* The phrase "sole discretion" does not appear in the cited provision, and the right to "suspend or revoke" access may also be subject to

	other contractual provisions—including notice requirements, implied covenants of good faith and fair dealing, and other terms of the parties' agreements—that Paragraph 679 omits. *Id.* Further, Mr. Melton had not seen that SOW and explained that he was "speculating and giving you generic company policy." Books Ex. 36 (L. Melton 30(b)(6) Tr.) at 94:22–96:4.

E. Microsoft Encouraged OpenAI's Data Practices, Embedded Its Engineers at OpenAI, and Obtained OpenAI's Technology

680. [redacted]	680. **Undisputed – Immaterial.** Microsoft responds that the statements in Paragraph 680 are immaterial because they do not affect the outcome of the suit under the governing law. Internal communications speculating about what Microsoft needed to do are not facts and immaterial to the issues before the Court. To the extent Plaintiffs purport to establish Microsoft's knowledge of the contents of OpenAI's training datasets, Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in Plaintiffs' summary judgment motion. OpenAI defers to Microsoft's response.
681. Microsoft placed engineers onsite at OpenAI. Nelson Decl. Ex. 59(a), at 99:18–100:25; Nelson Decl. Ex. 16 at 31:3–15. [redacted]	681. **Disputed in Part.** Defendants respond as follows: Defendants do not dispute that Exhibit 59(a) and Exhibit 16 contains the quoted language. But Defendants dispute any implication in the lead-in sentence to Paragraph 681 that the onsite engineer program was Microsoft's

[redacted]	initiative or something Microsoft controlled (*i.e.*, that "Microsoft *placed* engineers onsite at OpenAI")—because this statement mischaracterizes the nature of the arrangement. Ms. Templeton's testimony cited in Paragraph 681, establishes that: [redacted] C. Blythe Opp. Decl. Ex. 54 (D. Templeton 30(b)(6) Tr. (Jan. 20, 2026)) at 99:18-100:25; *see also* Books Ex. 146 at MSFT_AI_MDL_000000598-600 [redacted]
682. [redacted]	682. **Disputed in Part.** Defendants do not dispute that the cited evidence contains the quoted language. But Defendants dispute any implication in Paragraph 682 that Microsoft's onsite engineers at OpenAI could access OpenAI's training data. Microsoft's corporate designee testified that [redacted] C. Blythe Opp. Decl.

	Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 282:19–23; 267:20–25. OpenAI's corporate designee likewise testified that [redacted] C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 168:1–11.
683. [redacted]	683. **Disputed in Part – Not Supported by Cited Evidence; Incomplete.** Microsoft does not dispute that the cited evidence contains the quoted language. Microsoft disputes the characterization in Paragraph 683 that Mr. Zoph was designated to testify "on Microsoft's role in training" because it is not supported by the cited evidence. Mr. Zoph testified that he was "the *post*-training representative for OpenAI" and designated to speak on a variety of topics related to that, none which was "Microsoft's role in training." C. Blythe Opp. Decl. Ex. 66 (B. Zoph 30(b)(6) Tr.) at 13:6–15:21. Microsoft further disputes any implication in Paragraph 683 that Microsoft's onsite engineers or any other Microsoft employees played any role in OpenAI's training of GPT models because it is not supported by the cited evidence. Mr. Zoph's quoted testimony was in response to the question [redacted] and therefore does not support that Microsoft engineers had any role in

	training. C. Blythe Opp. Decl. Ex. 66 (B. Zoph 30(b)(6) Tr.) at 212:18–213:15. Mr. Zoph further testified that [redacted] *Id.* at 213:12–13. Further, Microsoft's corporate designee testified that [redacted] C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 282:19–23; 267:20–25. OpenAI's corporate designee likewise testified that [redacted] C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 168:1–11. OpenAI defers to Microsoft's response.
684. Under the Second Amended JDCA, Microsoft could embed up to twenty engineers during the first three years and thereafter the greater of twenty engineers or ten percent of OpenAI's research engineering personnel. Nelson Decl. Ex. 146 at -599. The agreement provides that onsite engineers "will take business direction from OpenAI with respect to prioritization of such Onsite Engineers' time and work tasks involving OpenAI," and "dedicate the overwhelming majority of their time as Microsoft employees (e.g., 70% - 85% of their full-time efforts) to performing as Onsite Engineers." *Id.*	684. **Undisputed.**

685. OpenAI gave Microsoft technical staff access to its training and evaluation code and to [redacted]. Nelson Decl. Ex. 168; Nelson Decl. Ex. 32 at 70:2–16. [redacted] Nelson Decl. Ex. 153. [redacted] *Id.* at -904.	685. **Disputed in Part – Not Supported by Cited Evidence; Incomplete.** Defendants do not dispute that OpenAI provided Microsoft with access to certain code, model weights, and APIs pursuant to the parties' contractual arrangements. However, disputed to the extent Plaintiffs' characterization is misleading as to the nature of these arrangements. The sharing described in Paragraph 685 occurred under the contractual framework of the parties' Joint Development and Collaboration Agreement ("JDCA"), which contemplated technology transfer between the parties—not as ad hoc disclosures by OpenAI. Defendants further dispute any implication in Paragraph 685 that OpenAI gave all Microsoft technical staff wholesale access to all of its training and evaluation code because it is not supported by the cited evidence. Mr. Madan testified that a limited number of "embedded engineers had limited access" to "training source code" pursuant to the limitations in the JDCA. C. Blythe Opp. Decl. Ex. 86 (U. Madan 30(b)(1) Tr.) at 70:1-25. As to the second and third sentences, the characterization that OpenAI shared access "following outreach from Nadella to Altman" is disputed because it implies that OpenAI was responding to pressure, when the exchange in fact reflects mutual collaboration under an existing contractual framework. The cited email shows that [redacted]

	[redacted] Books Ex. 153 at MSFT_AI_MDL_000006903. Further, the quoted statement attributed to Mr. Nadella regarding [redacted] *See* Books Ex. 153 at MSFT_AI_MDL_000006904.
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.	686. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Exhibit 40 contains the quoted deposition testimony. But, pursuant to Fed. R. Evid. 106, Mr. Nadella's full testimony includes how Microsoft's purpose for pursuing a partnership with OpenAI was informed by how "OpenAI came to the realization that in order for them to continue to be a lab that is doing leading work, they needed to fund a lot of compute and therefore they needed to create a for-profit entity and find a commercial partner who was willing to take risks, at which we took." *See* Books Ex. 40 (S. Nadella 30(b)(1) Tr.) 209:24-210:10. Microsoft did pursue a commercial relationship in its partnership with OpenAI, but in service of working collaboratively together to bring LLM technology to the general public and increasing competition. *See, e.g.*, Books Ex. 40 (S. Nadella 30(b)(1) Tr.) at 250:16-251:24 (describing the partnership as Microsoft's "strategic intent . . . to support AI companies in that time. . . . [O]ur entire goal was to sponsor anyone with ambition to compete at the frontier of AI. And that's what

	we were pushing ourselves to figure out a commercial construct that would allow us to do that. . . . In fact, that's sort of what attracted me to OpenAI is that they had, in some sense, the same ambition that we had, which is that we want to build great AI.").
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.	687. **Disputed in Part.** Microsoft does not dispute that in September 2020, Microsoft executives discussed internally how OpenAI's technology might be used in Microsoft's products, or that certain Microsoft engineering teams expressed a desire for access to GPT-3 training data [redacted] Books Ex. 154 at MSFT_AI_MDL_000007759. But Microsoft disputes the characterization in Paragraph 687 that Microsoft "need[ed]" to obtain the GPT-3 training data. In the same email chain, Microsoft's Chief Technology Officer, Kevin Scott, responded that he [redacted] *Id.* OpenAI defers to Microsoft's response.

F. Microsoft Could Control OpenAI's Training Data

688. In the Second Amended JDCA, OpenAI agreed not to, "without Microsoft's advance written agreement," knowingly "provide Microsoft any Model trained in a manner or using data that would preclude or limit the Commercial Use of such Model." Nelson Decl. Ex. 146 at -590.	688. **Disputed in Part – Incomplete.** Defendants do not dispute that Ex. 146(b) contains the quoted language. Pursuant to Fed. R. Evid. 106, Defendants respond that after the quoted language, the relevant 2023 JDCA provision states that "The foregoing does not apply to Models trained solely for research purposes" and that "OpCo and OpenAI may train all Models using data selected solely by OpCo and OpenAI." Books Ex. 146(b) at MSFT_AI_MDL_000000590.
689. Under the Second Amended JDCA, Microsoft could require OpenAI to respond to "requests to provide additional detail or answer specific questions regarding Models . . . (e.g., specifically what data was used to train any given Model)." Nelson Decl. Ex. 146 at -586; *see also* Nelson Decl. Ex. 40 at 59:15–60:11 (acknowledging Microsoft's "at least monthly opportunity to ask OpenAI about the data it trained on" via a "governing board").	689. **Undisputed.**
690. If Microsoft had "concerns regarding the use of any particular data to train a Model," OpenAI would "promptly retrain that Model using solely data approved by Microsoft." Nelson Decl. Ex. 146 at -590; *see also* Nelson Decl. Ex. 40 at 74:13–75:9.	690. **Disputed in Part – Incomplete.** Defendants do not dispute that Exhibit 146 contains the quoted language. But Defendants dispute the characterization in Paragraph 690 to the extent it suggests that the provision gave Microsoft control over OpenAI's selection of training data. The provision begins: "The primary purpose of this Agreement is to provide for the Commercial Use of Licensed Products powered by Licensed IP." Books Ex. 146(b) at –MSFT_AI_MDL_000000590.

	And it provides that, "[s]ubject to the foregoing, OpCo and OpenAI may train all Models using data selected solely by OpCo and OpenAI[.]" *Id.*
691. Microsoft's 30(b)(6) witness admitted that, despite concerns over copyright issues, Microsoft "instructed" its teams not to ask about OpenAI's training data. Nelson Decl. Ex. 28 at 246:4–247:6, 251:2–13.	691. **Disputed in Part – Incomplete.** Defendants respond as follows: Defendants do not dispute that Exhibit 28 contains the quoted language. But Defendants dispute the characterization of Paragraph 691 that Microsoft limited questions "despite concerns over copyright issues." Pursuant to Fed. R. Evid. 106, when asked why Microsoft gave that instruction, Mr. Huynh explained that "[t]he main one was really around resources from OpenAI," because "it takes time for somebody on OpenAI to answer specific questions," and "part of the partnership was around making the partnership as efficient as possible." C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 247:12–16. Mr. Huynh testified separately regarding: (1) awareness of copyright concerns; and (2) instructions to certain Microsoft teams not to ask about OpenAI's training data. Defendants dispute this fact as stated to the extent the SUF's use of the word "despite" implies a causal connection between the copyright concerns and the instruction not to ask about training data. The cited testimony at Books Ex. 28 at 246:4–247:6 and 251:2–13 addresses these subjects separately and does not establish that Microsoft instructed its teams not to ask about training data because of, or notwithstanding, copyright concerns. The SUF draws an

	inference not supported by the cited evidence. C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr. (Jan. 27, 2026)) at 246:4–247:6 and 251:2–13.
692. When Microsoft's CEO Satya Nadella was asked whether he would have invoked the contractual provision requiring OpenAI to retrain its models if he "had been made aware that OpenAI had scraped and trained on information that was behind a paywall," he responded: "I would." Nelson Decl. Ex. 40 at 76:20–77:2.	692. **Undisputed – Immaterial; Not Supported by Cited Evidence.** Microsoft does not dispute that this question and answer appear in Mr. Nadella's deposition but Microsoft responds that the hypothetical scenario described in Paragraph 692 is immaterial because it does not affect the outcome of the suit under the governing law. Microsoft further responds that Paragraph 692 is not supported by admissible evidence, as required by Local Rule 56.1. S.D.N.Y. Local Rule 56.1(d). The testimony is inadmissible because the question was an improper hypothetical, was incomplete, and called for speculation, including speculation about what was legally appropriate under the parties' agreements. In light of the objectionable question (all objections as to form were reserved), any testimony given in response is inadmissible. OpenAI defers to Microsoft's response.
693. When asked if he would have invoked the same provision if he "had learned that…OpenAI had trained its models on pirated works," Mr. Nadella answered: "Yeah, I mean, our entire principle was to do what is legally allowed and what is the industry norm and so yeah, I would have wanted to stick to that." *Id.* at 76:11–19.	693. **Undisputed – Immaterial; Not Supported by Cited Evidence.** Microsoft does not dispute that this question and answer appear in Mr. Nadella's deposition but Microsoft responds that the hypothetical scenario described in Paragraph 693 is immaterial because it does not affect the outcome of the suit under the governing law. Microsoft further responds that Paragraph 693 is not supported by admissible evidence, as

	required by Local Rule 56.1. S.D.N.Y. Local Rule 56.1(d). The testimony is inadmissible because the question was an improper hypothetical, was incomplete, and called for speculation, including speculation about what was legally appropriate under the parties' agreements. In light of the objectionable question (all objections as to form were reserved), any testimony given in response is inadmissible. OpenAI defers to Microsoft's response.
694. Microsoft had the unambiguous right to download model weights and code used by OpenAI. Nelson Decl. Ex. 60 at 236:6–16 ("[T]he JDCA is unambiguous that [Microsoft] have the rights to download code and it's unambiguous that [Microsoft] have the right to download model weight.").	694. **Undisputed.**
695. Microsoft could ask OpenAI what data OpenAI trained on, and likely did so. Nelson Decl. Ex. 40 at 60:7–61:10.	695. **Disputed in Part.** Microsoft does not dispute that Mr. Nadella testified that the 2023 JDCA provided a mechanism for Microsoft to ask OpenAI high-level questions about its training data. Books Ex. 40 (S. Nadella 30(b)(1) Tr. (Feb. 4, 2026)) at 60:7-11. But Microsoft disputes any implication in Paragraph 695 that doing so gave Microsoft knowledge of the specific data on which OpenAI trained its LLMs. In the same answer, Mr. Nadella explained that Microsoft asked "because at some fundamental level, we wanted to make our post training more effective," and that "this has been at least a source of a challenge for us. Even though we have IP, we don't have access to all the training data because of two reasons: One

	is how exactly is training data used in the training process is always opaque by definition, [redacted] Books Ex. 40 (S. Nadella 30(b)(1) Tr. (Feb. 4, 2026)) at 61:5-19. OpenAI defers to Microsoft's response.
696. Microsoft never exercised its contractual right to require OpenAI to retrain models using data Microsoft approved. Nelson Decl. Ex. 60 at 126:20–22 (Q: "Did Microsoft ever execute on its right to require OpenAI to retrain its model?" A: "No").	696. **Undisputed.**
697. Under the JDCA, Microsoft had the right to terminate its agreement with OpenAI if OpenAI breached the agreement, including by breaching intellectual property rights. Nelson Decl. Ex. 40 at 41:14–42:9; Nelson Decl. Ex. 145 at -364.	697. **Disputed in Part.** Defendants do not dispute that the cited JDCA provision in Ex. 145 permits termination for an uncured material breach following specified procedures. But Defendants dispute the implication that the termination right was uniquely held by Microsoft. The cited provision is mutual and states that "if the other party materially breaches this Agreement and fails to cure such breach within sixty (60) days after written notice thereof, a party may terminate on sixty (60) days' notice." Nelson Decl. Ex. 145 at MSFT_AI_MDL_000000364.
698. OpenAI was contractually required to "exclusively use Microsoft Azure and other relevant Microsoft supercomputing services for all AI training	698. **Undisputed.** Defendants do not dispute that OpenAI was to use Microsoft's Azure Services for AI training, inference, deployment, and commercial workloads throughout the term of the Amended JDCA. However,

inferenced deployment and commercial workloads." Nelson Decl. Ex. 60 at 337:10–21; Nelson Decl. Ex. 363 at -432.	Paragraph 698 cites the wrong document for this fact. Nelson Decl. Ex. 363 is a draft term sheet, not an executed JDCA. Although Mr. Tinter testified that he believed the provision was included in the final terms, the document he reviewed was likewise a draft. Books Ex. 60 at 337:18–21.
699. Microsoft determined in its sole discretion which Supercomputer Technology would be made available to OpenAI, including the method by which it would be made available and the requirements for granting access. Nelson Decl. Ex. 60 at 249:15–25, 270:20–271:10; Nelson Decl. Ex. 164 at -502 (Statement of Work No. 2 Supercomputer Tech Collateral Exchange with OpenAI); Nelson Decl. Ex. 149 at -491 (Amendment No. 3 to the Amended and Restated JDCA).	699. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Exhibits 164 and 149 contain the language paraphrased in Paragraph 699. But, pursuant to Fed. R. Evid. 106, the cited section of Exhibit 164 reveals that the "requirements for granting access" controlled only OpenAI personnel's access to Microsoft's technology. The provision covers "the requirements for granting access and permissions to access and use such Supercomputer Tech *to OpenAI and LP personnel*." Books Ex. 164 at MSFT_AI_MDL_000169502 (emphasis added). Microsoft's corporate designee testified that these conditions existed to "protect some of the technology that we brought to the collaboration." C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr.) at 274:12–14. When asked whether Microsoft ever "condition[ed] its access to this technology based on Microsoft's ability to review training data," Mr. Tinter answered: "No." *Id.* at 275:3–7. OpenAI defers to Microsoft's response.
700. Microsoft placed conditions on OpenAI's access to the technology Microsoft provided and retained the ability to terminate those services at various times, in some cases without cause. In practice, Microsoft never	700. **Disputed in Part.** Microsoft does not dispute that Exhibit 60 contains the noted language in Paragraph 700. But Microsoft does dispute the characterization in Paragraph 700 of the scope of those

conditioned its continued provision of that supercomputing technology and infrastructure on OpenAI's compliance with copyright law. Nelson Decl. Ex. 60 at 274:3–275:2.	conditions and termination rights afforded to Microsoft, including as being inclusive of IP violations. Mr. Tinter testified that the conditions served to "protect some of the technology that we brought to the collaboration," and that while [redacted] C. Blythe Opp. Decl. Ex. 8 (J. Tinter 30(b)(6) Tr. (Feb. 10, 2026)) at 274:12–14; 274:19–22. OpenAI defers to Microsoft's response.

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.	701. **Undisputed.** Defendants do not dispute that under the JDCA, OpenAI agreed to pay Microsoft twenty percent of net revenue for certain products. Books Ex. 146 at MSFT_AI_MDL_000000596.
702. Based on data produced for January 2023 through December 2024, Microsoft received [redacted] in net revenues from OpenAI related to [redacted] with no associated cost of goods sold such that the transferred revenue was also gross profit. Nelson Decl. Ex. 213 (tab "RevShare to MSFT"); Lasinski Decl. Ex. A at ¶ 167.	702. **Disputed in Part.** Microsoft does not dispute that Microsoft received [redacted] in worldwide net revenues from OpenAI related to [redacted] revenue from January 2023 through December 2024. But Microsoft disputes that there are no associated costs of goods sold related to this revenue. Microsoft's 30(b)(6) witness on financials testified that, while she "can't think of -- of a cost -- a cost of goods sold

	that would be specific to just receiving that rev share" that "there's other costs that Microsoft incurs for delivering Azure infrastructure to OpenAI so that OpenAI can generate their own revenue." *See* C. Blythe Opp. Decl. Ex. 67 (K. Griffith 30(b)(6) Tr.) at 55:24–56:10. Should Plaintiffs proceed to trial on any claims, Microsoft reserves the right to challenge any damages calculations put forward by Plaintiffs, including but not limited to challenging whether any such revenues are reasonably related to the alleged infringement. *On Davis v. Gap, Inc.*, 246 F.3d 152, 160 (2d Cir. 2001) ("the term 'gross revenue' under the statute means gross revenue reasonably related to the infringement, not unrelated revenues"). OpenAI defers to Microsoft's response.
703. For the period of July 2022 to December 2024, Microsoft generated [redacted] in revenues and $513.3 million in gross profits from Azure services purchased by OpenAI. Lasinski Decl. Ex. A at ¶ 170.	703. **Disputed – Not Supported by Cited Evidence.** Microsoft responds that the facts set forth in Paragraph 703 are disputed because they are not supported by the cited evidence. Microsoft disputes that it generated [redacted] in revenues and $513.3 million in gross profits from Azure services purchased by OpenAI for the period state. Those calculations represent worldwide revenues and gross profits from Azure services purchased by OpenAI and include revenues and profits for products not at issue in this litigation. ECF 1825-3 (M. Lasinski Decl. Ex. C) ¶ 53 (Mr. Lasinski acknowledging that his calculation of inferencing revenue is worldwide and includes revenue for models and products not at issue in the litigation). Should Plaintiffs

	proceed to trial on any claims, Microsoft reserves the right to challenge any damages calculations put forward by Plaintiffs, including but not limited to challenging whether any such revenues are reasonably related to the alleged infringement. *On Davis*, 246 F.3d at 160 ("the term 'gross revenue' under the statute means gross revenue reasonably related to the infringement, not unrelated revenues"). OpenAI defers to Microsoft's response.
704. Microsoft generated approximately [redacted] in revenue from providing Bing Web Search API services to OpenAI during April 2024 through December 2024, [redacted]. Nelson Decl. Ex. 212; Lasinski Decl. Ex. A at ¶ 172.	704. **Disputed – Not Supported by Cited Evidence.** Microsoft responds that the facts set forth in Paragraph 704 are disputed because they are not supported by the cited evidence. As to the first part of Paragraph 704, Microsoft generated approximately [redacted] in revenue from providing Bing Web Search API services to OpenAI during April 2024 through December 2024, not [redacted]. Books Ex. 212 (excerpt of MSFT_AI_MDL_002758533) at 3 (showing a grand total of [redacted]). Microsoft disputes any implication that this revenue relates to the Asserted Works. Should Plaintiffs proceed to trial on any claims, Microsoft reserves the right to challenge any damages calculations put forward by Plaintiffs, including but not limited to challenging whether any such revenues are reasonably related to the alleged infringement. *On Davis*, 246 F.3d at 160 ("the term 'gross revenue' under the statute means gross revenue reasonably related to the

	infringement, not unrelated revenues"). As to the second part of Paragraph 704, Microsoft disputes that this revenue is [redacted] C. Blythe Opp. Decl. Ex. 67 (K. Griffith 30(b)(6) Tr.) at 111:12–113:4. OpenAI defers to Microsoft's response.
705. [redacted]	705. **Undisputed – Incomplete.** Microsoft does not dispute that Exhibit 172 contains the quoted language. Pursuant to Fed. R. Evid. 106, Microsoft responds that Exhibit 172 was co-authored by Amy Hood and Kevin Scott. OpenAI defers to Microsoft's response.

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.	706. **Disputed in Part**. OpenAI responds as follows: Undisputed that Books Ex. 710 contains the quoted language. However, disputed that Plaintiffs' characterization of this statement is accurate or complete. Pursuant to Fed. R. Evid. 106, Defendants object because Plaintiffs' assertion omits critical context from the same email chain. In February 2019, David Lansky wrote, "In a nutshell, [redacted] to use as training data for our ML research, and [redacted] Books Ex. 710 at OPCO_MDL_004618324. Ms. Amodei's statement was made in response to a hypothetical question asking "Will the research search ever be used [redacted] and her full response was: [redacted] *Id.* at OPCO_MDL_004618322 to OPCO_MDL_004618323. Ms. Amodei also stated in the same email that "We definitely don't want to replace authors." *Id.* at OPCO_MDL_004618323. Microsoft defers to OpenAI.

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.	707. **Disputed in Part**. OpenAI responds as follows: Undisputed that Books Ex. 283 contains the quoted language. However, pursuant to Fed. R. Evid. 106, Defendants object because Plaintiffs' characterization of the testimony is inaccurate and misleading. The cited testimony does not demonstrate that OpenAI ever co-created fiction. Plaintiffs mischaracterize the testimony to suggest a conclusion that the evidence does not support. Instead, Amodei responded to Plaintiffs' attorney's leading question on [redacted] *See* C. Blythe Opp. Decl. Ex. 33 (Dario Amodei 30(b)(1) Tr.) at 114:12-15. Microsoft defers to OpenAI.
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.*	708. **Disputed in Part**. OpenAI responds as follows: Undisputed that Exhibit 283 contains the quoted text. However, disputed that "Among the first goals OpenAI proposed to demonstrate the model's capabilities was writing a short story." The cited portion of the document, which appears to be a draft only, discusses the possibility of "fine-tuning" a model in the future on "small datasets to mimic the style of a specific writer or new genre of writing." The cited portion of the document does not state that among the first goals OpenAI proposed to demonstrate the model's capabilities was writing a short story. *See* Books Ex. 283 at OPCO_MDL_001687108. Microsoft defers to OpenAI.

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.").	709. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 709 are disputed because they are not supported by the cited evidence. Specifically, the cited document does not indicate that the evaluations were "designed to measure creative writing and narrative performance," nor does it compare evaluations between "models trained with the LibGen books" and "models trained on general web-crawl data." Rather, the document discusses results on "downstream tasks for different mixes of the new data." Books Ex. 299 at OPCO_MDL_001903547. Microsoft defers to OpenAI.
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.	710. **Undisputed.** Microsoft defers to OpenAI.
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.	711. **Disputed in Part**. OpenAI responds as follows: Undisputed that Nelson Ex. 235 contains the quoted language. However, the facts set forth in Paragraph 711 are disputed because they are not supported by the record. Specifically, the evidence cited is an internal Slack communication that does not discuss or establish any of Plaintiffs' characterizations regarding a "book-summarization model," a process that "included sending full-text copies to third-party contractors," or why any such model is one in dispute in this litigation. Rather, the

	purpose of the underlying project was [redacted] C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) at 121:23–122:8; *see also id.* at 126:22–127:23; C. Blythe Opp. Decl. Ex. 35 at OPCO_MDL_000000069-70; ECF 707 (Stein Opinion & Order re Models and MTS in Class). As part of that effort, Mr. Brockman discusses a summary and notes that the model being discussed "only gets the first 3rd of the book right now." Further, and as discussed in response to Fact 712, Plaintiffs' characterizations of, and inferences drawn from, Mr. Mayne's statement are disputed as they ignore other statements by Mr. Mayne fully explaining his views. Microsoft defers to OpenAI.
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.	712. **Disputed in Part**. OpenAI responds as follows: Undisputed that Books Ex. 241 contains the quoted language. However, Paragraph 712 is disputed because Plaintiffs' characterization of Mr. Mayne's statement is incomplete and misleading. Mr. Mayne's full statement is that he has "found the quality of fiction writing in both storytelling and structure improves if you provide it with a sample in the prompt"—that is, by providing a sample at the time of prompting the model, not by incorporating books into training data. Books Ex. 241 at

	OPCO_MDL_000211503. Mr. Mayne further explained that the model has an idea of "what" a story is from online discussions, but that "actually writing a story and getting beats is different than what you learn from literary analysis[,]" and that the model "would likely have no experience of their writing unless the dataset was carefully curated." *Id.* By omitting this context, Plaintiffs mischaracterize Mr. Mayne's observations as support for the proposition that OpenAI improved fiction quality by training on copyrighted books, when Mr. Mayne was in fact describing user-directed prompting techniques and acknowledging limitations in the model's familiarity with authors' actual writing. Microsoft defers to OpenAI.
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.	713. **Disputed in Part – Not Supported by Cited Evidence**. OpenAI responds as follows: Undisputed that Books Ex. 352 contains the quoted language. However, The the facts set forth in Paragraph 713 are disputed because they are not supported by the record. Specifically, the cited document does not indicate that OpenAI "trained its models on books to improve their performance on creative writing." Rather, the document states that "we wish to improve" various capabilities, including "creative writing," and proposes to "upweight datasets of books/ whatever creating writing we might have." Books Ex. 352 at OPCO_MDL_004740422. Plaintiffs' assertion that OpenAI "trained" its models on

	books for this purpose mischaracterizes the cited evidence, which reflects a proposal to explore certain training adjustments, not that OpenAI in fact trained its models on books to improve creative writing performance. The document further describes the effort as a "Proposal" to [redacted] and notes that the team was [redacted] Books Ex. 352 at OPCO_MDL_004740422. Microsoft defers to OpenAI.
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.	714. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 714 are disputed because they are not supported by the record. The cited evidence discusses Mr. Gogineni's work at OpenAI in 2025. It does not state that he was hired in 2022 to lead efforts to improve the writing quality of OpenAI's models. *See* C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 36:14–20; 38:22–39:3. Further, Plaintiffs' characterization of Mr. Gogineni's role is inaccurate or incomplete because the cited testimony does not support the assertion that he was hired "to lead" OpenAI's efforts to improve writing quality. See *id.* Microsoft defers to OpenAI.

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.	715. **Disputed in Part**. OpenAI responds as follows: Undisputed that Mr. Gogineni testified that he understood his presence on Twitter to have been part of the impetus for his recruitment to OpenAI. However, disputed to the extent that Paragraph 715 suggests that Mr. Gogineni's personal social media posts are statements by OpenAI. *See* C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 36:7–12. Microsoft defers to OpenAI.
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.	716. **Disputed in Part**. OpenAI responds as follows: Undisputed that Mr. Gogineni described himself as having "an undue amount of power over the english language." However, disputed that Mr. Gogineni stated that his "research mission" was to create a machine that would supplant human authors. Mr. Gogineni did not state that his research mission was "to create a machine that would supplant authors." Rather, in response to a post on X suggesting a "proposal for a new LLM creative writing benchmark, 'can it write the last two books of A Song of Ice And Fire,'" Mr. Gogineni wrote, "this has been my research mission." Mr. Gogineni testified that "I think it would be very unfair to characterize that as my research goal — having been just, frankly, inaccurate," and that "nothing I have to do in my day-to-day has anything to do with that and is about improving model writing quality overall across a broad set of capabilities." *See* C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 71:5–11. Mr. Gogineni further

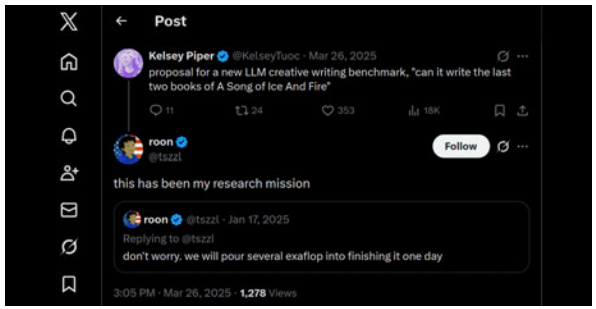
	explained that his tweets "represent fleeting thoughts" that "don't even represent necessarily things that I feel are true, even in the moment," and that they are "not meant to represent my official views, anyone else's official views, but generally fleeting thoughts in the moment." C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 69:17–23; 70:12–22. Mr. Gogineni also testified that his statement regarding having "an undue amount of power" over the English language was intended as a recruitment post, explaining: "I take this kind of tone in — in attempts to recruit, I think — exaggerating the importance of a product — or project, rather. And it can be valuable when recruiting. So it should be seen in that way." C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 41:11–17. Further disputed to the extent that Paragraph 716 suggests that Mr. Gogineni's personal social media posts are statements by OpenAI. See C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) 36:7–12. Microsoft defers to OpenAI.
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.	717. **Undisputed.** Microsoft defers to OpenAI
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	718. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited quotations are contained in the cited documents. However, disputed that Mr. Gogineni's personal social media posts are statements from

"astonished[,] more creative the genre, the better." Nelson Decl. Ex. 715; Nelson Decl. Ex. 98.	OpenAI. *See* Books Ex. 98; Books Ex. 715. These particular posts predate his employment with OpenAI. *See* Books Ex. 98; Books Ex. 715. Moreover, Exhibit 98 does not mention GPT-3. *See* Books Ex. 98. Microsoft defers to OpenAI.
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.*	719. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited document contains the cited quotations, and that it also states: "GPT-4 is more . . . collaborative than ever before." *See* Books Ex. 91 at 2. However, disputed that the cited document shows that OpenAI's executives publicized the models' creative writing improvements. *See id.* Microsoft defers to OpenAI.
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.	720. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited document contains the cited quotations. See Books Ex. 88. However, disputed under Fed. R. Evid. 106 to the extent that this fact suggests that models produce copies of training data. As the full quotation states: "It's becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distribution that don't matter, like what photos humans are likely to take or words humans commonly write down." *See* Books Ex. 88.

	Microsoft defers to OpenAI.
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.	721. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 721 are disputed because the cited evidence is does not support the claim. *See* Books Ex. 354 at OPCO_MDL_004760983. Specifically, the cited document is entitled "Writing Brainstorm Notes 3/28." *See id.* at OPCO_MDL_004760981. It indicates that Mr. Gogineni "is building a style specific RM [Reward Model]." *See id.* at OPCO_MDL_004760983. The cited quotation does not address pretraining. See *id.* Microsoft defers to OpenAI.
722. [redacted] while paying nothing for the books it took from LibGen. Nelson Decl. Ex. 288.	722. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed that Ex. 288 is relevant to LibGen. The cited exhibit does not mention LibGen. Disputed that the cited evidence has anything to do with pretraining or data used for pretraining. As Mr. Gogineni testified, the discussion in this document is "in the context of specific posttraining evaluations." C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 85:19-91:6. Microsoft defers to OpenAI.
723. OpenAI described [redacted]	723. **Disputed in Part.** OpenAI responds as follows: Undisputed that Nelson Decl. Ex. 354 contains the

[redacted] and as the [redacted] Nelson Decl. Ex. 354 at -981–82.	quoted statements. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those statements suggest that "writing" refers primarily to generating creative works or books. Ex. 354 identifies a broad range of writing uses, including communications such as emails, texts, and chats; marketing materials such as blogs, tweets, and LinkedIn posts; work writing such as reports, proposals, memos, and briefs; personal writing; academic writing; and creative writing such as stories and poems. *See* Books Ex. 354 at OPCO_MDL_004760981–82, OPCO_MDL_004760985. Microsoft defers to OpenAI.
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.	724. **Disputed in Part.** OpenAI responds as follows: Undisputed that Ex. 354 contains the quoted statements. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those statements suggest that "writing" refers primarily to generating creative works or books. Books Ex. 354 identifies a broad range of writing uses, including communications such as emails, texts, and chats; marketing materials such as blogs, tweets, and LinkedIn posts; work writing such as reports, proposals, memos, and briefs; personal writing; academic writing; and creative writing such as stories and poems. Books Ex. 354 at OPCO_MDL_004760981-85. Further disputed to the extent Plaintiffs suggest that OpenAI's stated goal or "north star" was only to increase the number of users generating

	written content. Rather, Books Ex. 354 states that the "Goal here is to identify opportunities to make writing dramatically better." *Id.* at OPCO_MDL_004760982. Ex. 354 also includes a note from "Peter" stating that "In my mind the north star is the Writing WAU - and the way we get there is by grounding in what people are looking to do, and the way to get there is primarily through actually improving the skill itself." *Id.* at OPCO_MDL_004760985. Microsoft defers to OpenAI.
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.	725. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quotations are contained in the cited document. *See* Books Ex. 130. However, disputed because the facts stated in Paragraph 725 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. As Mr. Altman states, "not sure yet how/when it will get released[.]" *Id.* Microsoft defers to OpenAI.
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.	726. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the quotations are contained in the cited document. Disputed to the extent this fact suggests that the phrase "a democracy of ghosts" is copyrightable or that this phrase as used in the cited evidence as copied from the book Pnin. *See* Books Ex. 108; Books Ex. 130. The cited evidence does not support that suggestion. *See* Books Ex. 108; Books Ex. 130.

	Microsoft defers to OpenAI.
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.  	727. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited evidence contains the quoted statements. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those statements suggest that writing the last two books of *A Song of Ice and Fire* was Gogineni's actual research mission at OpenAI. Gogineni testified that "I think it would be very unfair to characterize that as my research goal -- having been just, frankly, inaccurate." C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 71:9–11. He further testified that "I would say that nothing I have to do in my day-to-day has anything to do with that and is about improving model writing quality overall across a broad set of capabilities." *Id.* at 71:17–20. Gogineni explained that his tweets "represent fleeting thoughts. They don't even represent necessarily things that I feel are true, even in the moment. But the platform encourages incendiary and Overton window-stretching -- getting people to think about AI progress and dramatizing it" and that his tweets are "not meant to represent my official views, anyone else's official views, but generally fleeting thoughts in the moment." *Id.* at 69:18–23, 70:16–18. Microsoft defers to OpenAI.
728. Gogineni also posted that OpenAI would "pour several exaflop [of computer power] into finishing it	728. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited evidence contains the

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.	quoted statements. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those statements suggest that writing the last two books of *A Song of Ice and Fire* was Gogineni's actual research mission at OpenAI. Gogineni testified that "I think it would be very unfair to characterize that as my research goal -- having been just, frankly, inaccurate." C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 71:9–11. Gogineni also testified that "I would say that nothing I have to do in my day-to-day has anything to do with that and is about improving model writing quality overall across a broad set of capabilities." *Id.* at 71:17–20. Gogineni explained that his tweets "represent fleeting thoughts. They don't even represent necessarily things that I feel are true, even in the moment. But the platform encourages incendiary and Overton window-stretching -- getting people to think about AI progress and dramatizing it," and that his tweets are "not meant to represent my official views, anyone else's official views, but generally fleeting thoughts in the moment." *Id.* at 69:18–23, 70:16–18. Gogineni further explained that "I don't think a 300,000-token-length transformer could at all autocomplete the Song of Ice and Fire. It doesn't make sense." *Id.* at 70:19–22. Microsoft defers to OpenAI.
729. Gogineni was aware that "[G]eorge's lawyers might have a bone to pick" with his research mission. Nelson Decl. Ex. 747.	729. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 747 contains the quoted

	language. Disputed that the cited evidence says anything about Mr. Gogineni's research mission or that it shows Mr. Gogineni was actually "aware" of any statement by or belief of "george's lawyers." See Books Ex. 747; C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 73:17–74:6. Microsoft defers to OpenAI.
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.	730. **Disputed in Part.** OpenAI responds as follows: Undisputed that Brockman purchased books from Amazon. See Books Ex. 364 at OPCO_MDL_PD_000006808. However, disputed that the books were "written by GPT-2." Brockman testified "I believe I may have purchased a book -- possible multiple, I don't know -- where the author said that they wrote together with GPT-2." *See* C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 204:22–205:2. Disputed that cited document shows that OpenAI had a "library." *See* Books Ex. 719; C. Blythe Opp. Decl. Ex. 2 (G. Brockman 30(b)(1) Tr.) at 204:19–205:2. Brockman instead described adding the books to the "office bookshelf." *See* Books Ex. 719. Microsoft defers to OpenAI.
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	731. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** OpenAI responds as follows: Undisputed that the cited exhibits contain the quoted language. Disputed that the facts they are supported by the cited evidence. Disputed that Plaintiffs' characterization that Gogineni

sympathetic" and instead viewed them as "acceptable economic disruption." Nelson Decl. Ex. 723.	"recognized that the model had lifted phrases from Nabokov" accurately reflects the cited evidence. Rather, when discussing a model not at issue in this case, Gogineni stated that the phrase "a democracy of ghosts" was "originated by Nabokov," but characterized the model's use of the phrase as an "original usage." Books Ex. 99. Further disputed that Plaintiffs accurately characterize Gogineni as having been "aware of authors' complaints that the 'datasets are stolen.'" The cited post, which predates Gogineni's employment at OpenAI, states that Gogineni had been trying to understand artists' complaints about AI art and characterized those complaints as including that "datasets are stolen, authors underpaid" and that artists saw some junior colleagues had lost "a gig or two to the AI." Books Ex. 723; *see* C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 10:23–24. Further, the facts set forth in Paragraph 731 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, to the extent Plaintiffs rely on the cited post to establish the truth of third parties' assertions that datasets were "stolen" or that authors or artists lost work to AI-generated competition; those assertions are hearsay offered for their truth. Fed. R. Evid. 802; Books Ex. 723. Microsoft defers to OpenAI.

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.	732. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited document contains the cited language. However, Plaintiffs' characterization is inaccurate or incomplete because as Mr. Gogineni testified, this was "clearly meant to be an absurdist tweet." *See* C. Blythe Opp. Decl. Ex. 68 (T. Gogineni 30(b)(1) Tr.) at 98:19-99:14. Microsoft defers to OpenAI.
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.	733. **Disputed in Part.** OpenAI responds as follows: Undisputed that Mr. Gogineni wrote that it would soon become "impossible to tell" whether material was AI-generated. However, disputed that Plaintiffs' characterization of these posts is accurate or complete. These posts from Mr. Gogineni's personal X account refer to a specific article and state "this website is becoming full of gpt-generated writing. obviously the well informed aristocracy of reading skill can see and feel that something's off for now, but the undertold story is that it'll pretty soon become impossible to tell." Books Ex. 722 at 1. Plaintiffs' selective quotation omits this context and mischaracterizes Mr. Gogineni's statements. Microsoft defers to OpenAI.
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	734. **Disputed in Part – Incomplete, Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Books Ex. 748 contains the quoted language. However, disputed facts set forth in Paragraph

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.	734 are disputed because they are not supported by the cited evidence. Specifically, Mr. Gogineni did not write that the models' writing capabilities depended entirely on "vast lineage of human mastery recorded in bytes, without which models could not do anything great." Rather, pursuant to Fed. R. Evid. 106, he wrote on his personal X account that "the glory of any current and future achievements of the models is collectively owned by the vast lineage of human mastery recorded in bytes, without which models could not do anything great." Books Ex. 748. Microsoft defers to OpenAI.
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.	735. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 735 are disputed because it is not supported by the cited evidence. Specifically, Mr. Gogineni did not write that "frontier large language models" are "channeling the accumulated skills of every brilliant human that recorded themselves before you." Rather, in a post on his personal X account, he wrote: "using a frontier ai is less like summoning an alien intelligence and more like going avatar state and channeling the accumulated skills of every brilliant human that recorded themselves before you. the heritage and future of humankind." Books Ex. 749. Microsoft defers to OpenAI.
736. Users have in fact employed the Accused Models to write sequels to Class Plaintiffs' works. Searching	736. **Disputed – Not Supported by Cited Evidence.** Defendants respond as follows: The facts set

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.*	forth in Paragraph 736 are disputed because they are not supported by the cited evidence. Specifically, the cited evidence does not show that "[u]sers have in fact employed the Accused Models to write sequels to Book Plaintiffs' works." A user asking ChatGPT to speculate about what might happen in an unpublished book is not "writ[ing] a sequel." *See also* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 23, 25-26 (demonstrating that verbatim 25-word excerpts of Martin's work appeared in 8 out of 95M messages (0.000008%)); S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 48-50 (demonstrating that Shan's methodology lacks error-rate validation and produces false positives).
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.	737. **Disputed – Not Supported by Cited Evidence.** See Defendants' response to Paragraph 736.
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.*	738. **Disputed in Part – Incomplete.** Defendants do not dispute that the model provided a 501-word response. But, pursuant to Fed. R. Evid. 106, the model's response was described as "a speculative one-page summary of what might happen in the sixth book of 'A Song of Ice and Fire,' titled 'The Winds of Winter.'" ECF 1814-4 (S. Shan Decl. Ex. D) ¶ 72 n.119.
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,'"	739. **Disputed in Part – Incomplete.** Defendants respond as follows: Undisputed that the cited evidence contains the cited quotations. But, pursuant to

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.*	Fed. R. Evid. 106, , disputed that "the model returned a full opening chapter." Rather, the model provided a 320-word response. ECF 1814-4 (S. Shan Decl. Ex. D) ¶ 72; *see also* ECF 1750 (T. Berg-Kirkpatrick Decl.) ¶¶ 23, 25-26 (demonstrating that verbatim 25-word excerpts of Martin's work appeared in 8 out of 95M messages (0.000008%)); S. Krishnan Opp. Decl. (Books) Ex. A ¶¶ 48-50 (demonstrating that Shan's methodology lacks error-rate validation and produces false positives).
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.	740. **Disputed in Part.** OpenAI responds as follows: Undisputed that the cited webpage contains the cited quotations. However, disputed to the extent this fact suggests that OpenAI markets ChatGPT to do creative writing instead of having humans do it. Rather, the document discusses ways ChatGPT can be used to help humans with their writing. Books Ex. 90. Microsoft defers to OpenAI.
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.	741. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Ex. 112 contains the quoted language. But, pursuant to Fed. R. Evid. 106, the same page states that "[w]hile Copilot is a valuable tool, it doesn't replace human storytelling. Your voice, originality, and emotional depth are irreplaceable," and directs writers to "ensure that their work reflects their own voice and vision, using AI-generated suggestions as inspiration rather than direct replacements." Books Ex. 112 at 3. It further states that "[r]esponsible use

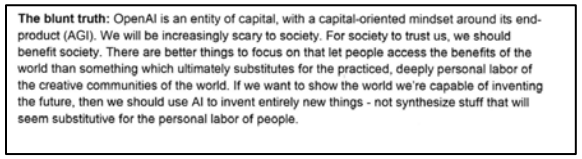
	also means being mindful of copyright, avoiding replication of existing works, and crediting sources where appropriate." *Id.* OpenAI defers to Microsoft's response.
742. A page titled "Write your book with AI help" states that "[w]hether you want to publish a sweeping epic or share a cozy children's 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.	742. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Ex. 114 contains the quoted language. But, pursuant to Fed. R. Evid. 106, the full exhibit describes Copilot as an "assistant" that helps a person "share [their] tales and stories" and "bring their ideas to life." Books Ex. 114 at 1. It explains that during "the drafting stage" Copilot "can provide real-time writing assistance" by giving "feedback for consistent tone, sentence structure, and stylistic improvements," and its sample prompts ask for feedback on the author's own work, such as "Can you give me feedback on this outline?" *Id.* at 2–3. OpenAI defers to Microsoft's response.
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.	743. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Ex. 113 contains the quoted language. But, pursuant to Fed. R. Evid. 106, the full exhibit describes creative writing as "a deeply personal and imaginative process" and Copilot as an "AI writing assistant" that helps writers "spark [their] creative inspiration for the stories [they] were meant to tell," by "offering prompts, alternative phrasings, and thematic suggestions tailored to your writing style." Books Ex. 113 at

	1–2. The prompt quoted in Paragraph 743 asks for tropes the author "could weave into the narrative." *Id.* at 2. OpenAI defers to Microsoft's response.
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.	744. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Ex. 111 contains the quoted language. But, pursuant to Fed. R. Evid. 106, the full exhibit advises writers to "[r]efine the text with your own tastes, experiences, and style," explains that "[e]very writer brings their own lived experiences to their text," and states that "anything that AI generates is best used as a launchpad of ideas, rather than a finished product." Books Ex. 111 at 2. OpenAI defers to Microsoft's response.
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.	745. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 745 are disputed because they are not supported by the cited evidence. Specifically, the cited evidence does not state that "training on books improves a model's ability to generate book-like text." Rather, it discusses the relationship between diversity of training data and the linguistic patterns that GPT 1 and 2 learned to generalize. *See* Books Ex. 751, ¶¶ 333–35. Microsoft defers to OpenAI.

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.	746. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the cited exhibit contains the quoted language. However, the facts set forth in Paragraph 746 are disputed because they are not supported by the cited evidence. The cited document is a deposition, not congressional testimony. During the deposition, Dr. Callison-Burch was asked to read excerpts of his congressional testimony into the record, including the following: "I downloaded 15,000 poems from the Internet and trained a system to write much better poetry than I ever could." C. Blythe Opp. Decl. Ex. 58 (C. Callison-Burch Tr. (Jul. 16, 2026)) at 488:8-11. Dr. Callison-Burch explained that this was a "research experiment that I was conducting to try to understand supervised fine-tuning in language models to see if a particular language model could be fine-tuned to act in a certain fashion." *Id.* at 489:1-6. He further testified that "I don't think it was particularly successful. The aesthetic value of the poetry was not great." *Id.* at 489:16-18. Further, the congressional testimony concerned a post-training technique known as supervised fine-tuning, not pretraining, as Paragraph 746 suggests. *Id.* at 490:9-15. As Dr. Callison-Burch stated during his deposition, "this technique is different than pretraining." *See id.* at 492:9-10. Microsoft defers to OpenAI.
747. Microsoft researcher Corbin Rosset testified that "[i]f you want to train a model to write poetry, you	747. **Disputed in Part – Incomplete.** Microsoft responds as follows: Microsoft does not dispute

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.	that Mr. Rosset gave the quoted testimony. But, pursuant to Fed. R. Evid. 106, Mr. Rosset was addressing the general "trade-off between quantity and quality" in training data, explaining that "[t]he dataset should be correct and clean and high-quality" and that "[t]here can be noise in the dataset." Books Ex. 48 (C. Rossett 30(b)(1) Tr.) at 119:5-6, 120:3. He agreed that a poetry-specific dataset would be appropriate "[i]f that was the task in question." *Id.* at 120:12-13. OpenAI defers to Microsoft's response.
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.	748. **Undisputed – Incomplete.** Microsoft responds as follows: Microsoft does not dispute that Dr. Lafferty's rebuttal report includes the language quoted in Paragraph 748. But, pursuant to Fed. R. Evid. 106, Microsoft incorporates its response regarding Dr. Lafferty's testimony cited in response to Paragraph 496. Further, Lafferty opines that "[a]t most … exposure to news-like or book-like text may improve a model's ability to generate news-like or book-like outputs. They do not show that news or books have unique functional importance to general LLM capability. In general, the distributional properties of LLM outputs are derived from the properties of the text that it is trained on. In order to be more generally useful, a model should be trained on a wide diversity of types of text. A model can learn from the factual information, current events, explanatory writing, multiple perspectives, and domain-specific language that can appear in many sources.

	No single alternative source would necessarily provide the same training value as a broad data mixture. The characteristics and attributes Plaintiffs' Experts identify are not unique to licensed news content or to any specific publisher's articles." Books Ex. 727 (J. Lafferty Rebuttal Rep.) ¶¶ 69-70. OpenAI defers to Microsoft's response.
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.	749. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 749 are disputed because they are not supported by the cited evidence. Disputed. The cited evidence does not say that Mr. Clark (or any other OpenAI employee) "knew that using copyrighted works without compensating their authors would harm those authors." Rather, it discusses Mr. Clark's summarization of the general concerns of some unidentified creators. The cited evidence does not discuss "copyrighted works." Microsoft defers to OpenAI.
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,	750. **Disputed in Part.** OpenAI responds as follows: Undisputed that Ex. 282 contains the cited statements. However, disputed that the selective quoting accurately reflects the cited evidence. The parts of Clark's statement that were removed are reproduced in bold below.

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.	"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 Al to invent entirely new things - not synthesize stuff that will seem substitutive for the personal labor of people." Books Ex. 282 at OPCO_MDL_001686929. The omitted language demonstrates that Mr. Clark was recommending a product strategy to earn public trust, not acknowledging that OpenAI was using copyrighted works without compensation or that doing so would harm authors. Microsoft defers to OpenAI.
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.	751. **Disputed in Part – Incomplete; Immaterial.** Microsoft responds as follows: Microsoft does not dispute that Brent Hecht prepared and presented a presentation in December 2022. Pursuant to Fed. R. Evid. 106, with regard to Exhibit 185 and Hecht's testimony relating to this exhibit, Hecht characterized the paper as a hypothesis and testified that "the jury is still out on a lot of this stuff." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 166:10–11; *see id.* at 165:9–15 ("It is again, we have

	to understand that as a hypothesis that, again, the evidence is still coming in.”). Microsoft further responds that the facts stated in Paragraph 751 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee’s personal views on any hypothetical effects of LLMs on authors have no bearing on any element of the claims or defenses in this case. Microsoft also disputes the facts set forth in Paragraph 751 for the following reasons: Hecht’s statements cannot be attributed to Microsoft. Hecht is a research academic employed at Microsoft to bring “divergent,” “asymmetric, futuristic, and academic points of view”; he does not set Microsoft’s core priorities or speak for Microsoft as to “his theoretical views on AI’s potential effect on content creators”; and Microsoft leadership “does not share all of [his] views;” and he is a “kind of intellectual provocateur.” ECF 1707 J. Usdan Decl. ¶ 39; C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 173:22-174:9. Hecht has himself disclaimed any authority to speak for Microsoft, testifying that his views on these subjects are “my view, not Microsoft’s view,” that “[he] has no idea if [they are] shared by the board of Microsoft,” and that “[he] can’t speak for the company on this stuff.” C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 319:18–20, 320:10–12, 321:4–5. He further described his views as “hypotheses,” cautioned that

	"the jury is still out on a lot of this stuff," and acknowledged that he is "not a lawyer." *Id.* at 59:11–60:8, 165:9–15, 166:10–11. Hecht's personal views therefore are not statements of Microsoft and do not reflect Microsoft's knowledge, policies, or conduct. OpenAI joins Microsoft's response. Hecht is a Microsoft researcher who testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802.
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.	752. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Exhibit 185 contains the quoted language. Pursuant to Fed. R. Evid. 106, with regard to Exhibit 185 and Hecht's testimony relating to this exhibit, Hecht characterized the paper as a hypothesis and testified that "the jury is still out on a lot of this stuff." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) 169:1-7 ("It is again, we have to understand that as a hypothesis that, again, the evidence is still coming in."); 166:2–11; *see id.* at 165:9–15. Microsoft further responds that the facts stated in Paragraph 752 are immaterial and therefore disputed because these facts do not affect the outcome of the suit

	under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on content creators have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response.
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.	753. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Books Ex. 185 contains the quoted language. Pursuant to Fed. R. Evid. 106, with regard to Exhibit 185, Hecht characterized the paper as a hypothesis and testified that "the jury is still out on a lot of this stuff." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 169:1-7 ("It is again, we have to understand that as a hypothesis that, again, the evidence is still coming in."); 166:2–11; *see id.* at 165:9–15. Microsoft further responds that the facts stated in Paragraph 753 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of content creators on LLM development have no bearing on any element of the claims or defenses in this case.

	Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response.
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.	754. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Ex. 185 contains the quoted language. Pursuant to Fed. R. Evid. 106, Exhibit 185 does not state that Microsoft would be seen as "the Napster of AI." It states that, "[e]ven if we were to assume the above scenarios do not fully play out," "[i]t is easy to imagine tenants like governments not wanting to inherit the reputational and legal risks of using products the public may begin to understand as the 'Napster of AI.'" Books Ex. 185 at MSFT_AI_MDL_000952900. Hecht also characterized Exhibit 185 as a hypothesis and testified that "the jury is still out on a lot of this stuff." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 169:1-7 ("It is again, we have to understand that as a hypothesis that, again, the evidence is still coming in."); 166:2–11; *see id.* at 165:9–15. Microsoft further responds that the facts stated in Paragraph 754 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on the company's public perception have no bearing on any element of the claims or defenses in this case.

	Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response.
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.	755. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Hecht characterized books as high-quality content. Pursuant to Fed. R. Evid. 106, Hecht's testimony was in the context of Exhibit 20 to his deposition, during which he testified that "[t]he specifics of where data is valuable and in which applications and stuff like this gets pretty gnarly as I was mentioning earlier, but in principle, high-quality data leads to high-quality outcomes by default." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 255:3-13. When asked whether published books are a source of high-quality content for LLMs, Hecht testified, "It depends on the book, but the same principle I mentioned before applies. With books in particular, there is a strong hypothesis that – the long form reasoning that books enable is important to – to language models." *Id.* at 259:17 – 260:2 Microsoft further responds that the facts stated in Paragraph 755 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on the hypothetical importance of particular data

	used for LLM development have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response. Hecht is a Microsoft researcher who testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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.	756. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Nelson Decl. Ex. 26 contains the quoted excerpts from Hecht's testimony. Pursuant to Fed. R. Evid. 106, Hecht's full testimony shows that his example of the "skeletons in the closet" was "what came out in the Meta lawsuit around how some of the data was obtained." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 59:20–24. When asked what he was referring to specifically, Mr. Hecht testified: "Again, not a lawyer, but something that was noticed by people in my research community for sure, the acquisition of books in a manner that has -- again, not a lawyer, but my interpretation has been interpreted by courts as -- as piracy and -- yeah, that would be one example of the skeletons in the closet." *Id.* at 60:2–8.

	Microsoft further responds that the facts stated in Paragraph 756 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal, lay views on any hypothetical legal interpretation about data acquisition for LLM development have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response, except to the extent any portion of it, including any testimony quoted under Fed. R. Evid. 106, could be read to characterize OpenAI, its models, or its conduct. OpenAI disputes any such characterization. Hecht is a Microsoft researcher who testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.
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	757. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Nelson Decl. Ex. 183 contains the quoted language. Pursuant to Fed. R.

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.	Evid. 106, when testifying about Exhibit 183, Hecht specifically testified: "Here is another disclaimer of me not being a lawyer. I know some of these claims are contested claims." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 106:10-12. Microsoft further responds that the facts stated in Paragraph 757 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on how hypothetically LLM development operates have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response. Hecht is a Microsoft researcher who testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.
758. The paper observed that this assumption is unlikely to hold and "may be very bad for Microsoft's	758. **Disputed in Part – Immaterial.** Microsoft does not dispute that Ex. 183 contains the quoted

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.	language. But, Microsoft responds that the facts stated in Paragraph 758 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on content creators have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 758 for the following reasons: The exhibit does not characterize the use of training data as theft; it predicts that "millions of people around the world will soon consider" it to be so, and that people "may also begin to go on 'data strikes.'" Books Ex. 183 at MSFT_AI_MDL_000949430. The document then states: "But enough about the risks … let's talk about the surprising opportunity: if we take the right actions now, all of this can be good news for Microsoft. The reason: We're really good at compliance." *Id.* OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
759. To address these issues, Hecht recommended that Microsoft focus on creating "compliant foundation	759. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Ex. 183 contains the

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.	quoted language. But, Microsoft responds that the facts stated in Paragraph 759 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on a recommended course of action for LLM development have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response.
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.	760. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Ex. 195 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 760 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on a recommended course of action for LLM development have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 760 for the following reasons: The "Entirely Reinvent" language appears as a proposed "**Grand Challenge** for MSR" (Microsoft Research). Books Ex. 195 at Slide 47 (emphasis added).

	OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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.	761. **Disputed in Part – Immaterial.** Microsoft does not dispute that Ex. 195 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 761 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on authors' perception of Microsoft have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 761 for the following reasons: When testifying about this slide, Hecht testified that the author's convention was a "hypothetic[al] – metaphorical authors convention to which I was referring." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 98:13-20. Microsoft also disputes Paragraph 761 to the extent it suggests that the slide in Books Ex. 195 describes Microsoft's conduct. Hecht testified that on this slide "I'm referring here to the field," and that it is "more of a critique of the direction of science,

	actually, than any -- any specific company's actions." *Id.* at 101:4–14. He further testified that "we're trying to milk too much out of an appendix slide" and that "I can't speak for sure about what I -- what I meant here." *Id.* at 102:7–10. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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.	762. **Disputed in Part – Immaterial.** Microsoft does not dispute that Nelson Decl. Ex. 26 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 762 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical understanding of LLM development on the part of book authors and newspapers have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
763. In a December 2023 email, Hecht wrote that "the current LLM paradigm could be on track for a doom	763. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books Ex.

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.	194 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 763 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on content creators and the AI industry have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 763 for the following reasons: Pursuant to Fed. R. Evid. 106, paragraph 763 omits that Hecht was relaying "[a] very important finding from the literature" that the paradigm "could be on track for a doom spiral," in an email thread concerning "BingChat and source references, UX considerations." Books Ex. 194 at MSFT_AI_MDL_001180173. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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,	764. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books Ex. 26 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 764 are immaterial and

the existence of the LLMs and the products based on them were at risk." Nelson Decl. Ex. 26 at 51:14–53:2.	therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on authors and newspapers have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 764 for the following reasons: Pursuant to Fed. R. Evid. 106, Hecht also testified that the "we" (which Paragraph 764 mischaracterizes as "[LLM producers"]) in "eating our own seed corn" "actually is everyone who is part of the production of AI systems which includes content producers," not merely LLM developers. C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 53:9–12. Further, Hecht described the doom loop as "a subject of scientific debate." *Id.* at 52:4–10. OpenAI joins Microsoft's response. Hecht is a Microsoft researcher who testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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	765. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books Ex. 26 contains the quoted language. But, Microsoft responds that

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).	the facts stated in Paragraph 765 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLMs on authors and LLM development at large have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 765 for the following reasons: Pursuant to Fed. R. Evid. 106, before reading Books Exhibit 202, Hecht cautioned that "this is not a peer-reviewed document" and if he "were writing a peer-reviewed document," that it would "require more revision." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 274:15-25. Hecht also described the doom loop as "a subject of scientific debate." *Id*. at 52:4–10. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12.
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	766. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Microsoft does not dispute that Ex. 179 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 766 are immaterial and therefore disputed because these facts

models can just learn from each other." Nelson Decl. Ex. 179 at -974.	do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical externalities from LLM development on generative AI and / or content valuation have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 766 because it is not supported by the cited evidence. Hecht did not state that the industry was "not paying sufficient attention" to anything; he shared an article and wrote that it "[h]ighlights some of the other externalities of the current status quo[.]" Books Ex. 179 at MSFT_AI_MDL_000533974. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.
767. Hecht testified that the risk that "producers of content such as newspapers and authors and books will	767. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Ex. 26 contains the

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.	quoted language. But, Microsoft responds that the facts stated in Paragraph 767 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLM development on authors and newspapers' businesses have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 767 for the following reasons: Hecht described such purported risk as "one hypothesized component of the paradox of reuse," and testified that "we have to understand that as a hypothesis that, again, the evidence is definitely still coming in," and that "the jury is still out on a lot of this stuff." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 166:10–11, 169:4–7, 292:11–13. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.

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.	768. **Disputed in Part – Immaterial.** Microsoft does not dispute that Ex. 26 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 768 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical effects of LLM development on authors' ability to sell their works have no bearing on any element of the claims or defenses in this case. Additionally, Hecht's statements are immaterial and cannot be attributed to Microsoft for the reasons stated in response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 768 for the following reasons: Hecht gave the answer in the context of what he described as "my view," which "I have no idea if it's shared by the board of Microsoft," and he testified immediately afterward that "there's going to be diverse opinions" and "I can't speak for the company on this stuff[.]" C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 320:10–12, 321:2–6. Further, Hecht described any concern as "one hypothesized component of the paradox of reuse," and testified that "we have to understand that as a hypothesis that, again, the evidence is definitely still coming in," and that "the jury is still out on a lot of this stuff." *Id.* at 166:10–11, 169:4–7, 292:11–13. OpenAI joins Microsoft's response. Hecht is a Microsoft researcher who testified unequivocally: "I don't

	speak for Microsoft. I definitely don't speak for OpenAI." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.
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.	769. **Disputed in Part – Improper Legal Argument; Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Ex. 342, which does not indicate the author of the document, contains the quoted language but the quote is misleading by omission. Pursuant to Fed. R. Evid. 106, the full sentence reads "We should have a plan for how we will deal with copyrighted data and a way for publishers/creators to get value, but ideally not make it part of the core training." Books Ex. 342 at OPCO_MDL_004619018. However, disputed that the cited documents support the fact that OpenAI executives recognized that rights holders should be compensated for the data used to train its models. The documents generally discuss potential methods to obtain additional data including the possibility of offering financial incentives. And the Court should disregard the disputed portion of Paragraph 769 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Further, the disputed proposition is not supported by the cited evidence.

	Microsoft defers to OpenAI.
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.	770. **Disputed in Part – Incomplete and Immaterial.** Microsoft does not dispute that Books Ex. 839 contains the quoted language. Pursuant to Fed R. Evid. 106, Vargas prefaced his answer, "Again, not a legal opinion obviously," and described his answers in this line of questioning as "my personal perspective." Book Ex. 839 (L. Vargas, Ph.D 30(b)(1) Tr.) at 90:25–91:11. When asked whether AI companies threaten the livelihood of authors by training on their works, Vargas testified: "I don't have enough understanding of the direct or indirect implications here to be able to elicit a valuable, a valuable perspective. I don't know." *Id.* at 91:24–92:8. Microsoft further responds that the facts stated in Paragraph 770 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical compensation practices for the use of books in LLM development have no bearing on any element of the claims or defenses in this case. Microsoft also disputes the characterization in Paragraph 770 that Microsoft witnesses gave "the same view" as purported statements made by OpenAI witnesses. Paragraph 770 cites only one Microsoft witness's personal,

	lay perspective, and that testimony does not compare his views to those of any OpenAI witness. OpenAI defers to Microsoft's response.
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.	771. **Disputed in Part – Immaterial; Incomplete.** Microsoft does not dispute that Books Ex. 65 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 771 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical compensation practice for creators of LLM training content have no bearing on any element of the claims or defenses in this case. Weyl "holds divergent, academic, and forward-looking views about how data ecosystems for AI should operate and is employed at Microsoft to bring these asymmetric, futuristic, and academic points of view." ECF 1707 (J. Usdan Decl.) ¶¶ 38-39. Microsoft also disputes the facts set forth in Paragraph 771 for the following reasons: Pursuant to Fed. R. Evid. 106, in Books Ex. 65, Weyl was discussing the "data dignity" framework he articulated in a 2018 article, which "primarily focused on a world that we expected to come into existence in the coming years, not on the technologies that were deployed at that time." Books Ex. 65 (E. Weyl 30(b)(1) Tr.) at 34:19-35:23. The "prototypical" content with which that framework was concerned "is, in principle,

	copyrightable but probably would not actually be subject to copyright monetization or other commercial intent and is … pretty sharply distinguished from the set of content that is produced by organizations that have commercial intent in producing copyrighted content." *Id.* at 38:17–39:1. And Weyl testified that the framework's "primary argument" was that such non-monetized content "may be even more important than is standard monetized copyrightable material as the basis for training these systems," and that his views regarding data creator compensation "wouldn't put special emphasis on" copyrighted content "and the fact that it is copyrighted." *Id.* at 39:2-6, 39:19-22. He also testified that "much less of the material and much less of the valuable material came from organized copyright holders than those copyright holders believed was the case, and that much more of it came from people who did not recognize themselves as engaging in economic activity." *Id.* at 136:19–137:6 *see also* ECF 1726 (C. Tucker Decl.) ¶¶ 140-145. OpenAI defers to Microsoft's response.
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.	772. **Disputed in Part – Immaterial; Incomplete.** Microsoft does not dispute that Ex. 26 contains the quoted language. But, Microsoft responds that the facts stated in Paragraph 772 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. A Microsoft employee's personal views on any hypothetical compensation practices for the use of books in LLM

	development have no bearing on any element of the claims or defenses in this case. Hecht is a research academic employed at Microsoft to bring "divergent," "asymmetric, futuristic, and academic points of view"; he does not set Microsoft's core priorities or speak for Microsoft as to "his theoretical views on AI's potential effect on content creators"; and Microsoft leadership "does not share all of [his] views." ECF 1707 (J. Usdan Decl.) ¶ 39. Microsoft also incorporates its response to Paragraph 751. Microsoft also disputes the facts set forth in Paragraph 772 for the following reasons: Pursuant to Fed. R. Evid. 106, the quoted language comes from Plaintiffs' counsel's question, to which Hecht responded: "That is included and goes well beyond authors, but, yeah." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 321:8–14. Mr. Hecht gave that answer in the context of a hypothetical in which "OpenAI is successful at its mission," "there is extensive job displacement," and he described it as "my view, not Microsoft's view." C. Blythe Opp. Decl. Ex. 69 (B. Hecht 30(b)(1) Tr.) at 319:13–321:1. He further testified that there are "diverse opinions" in the scientific community, he has "no idea if [his perspective is] shared by the board of Microsoft" and "I can't speak for the company on this stuff, and nor for OpenAI, certainly." *Id.* at 320:10–12, 321:2–6. OpenAI joins Microsoft's response. Hecht testified unequivocally: "I don't speak for Microsoft. I definitely

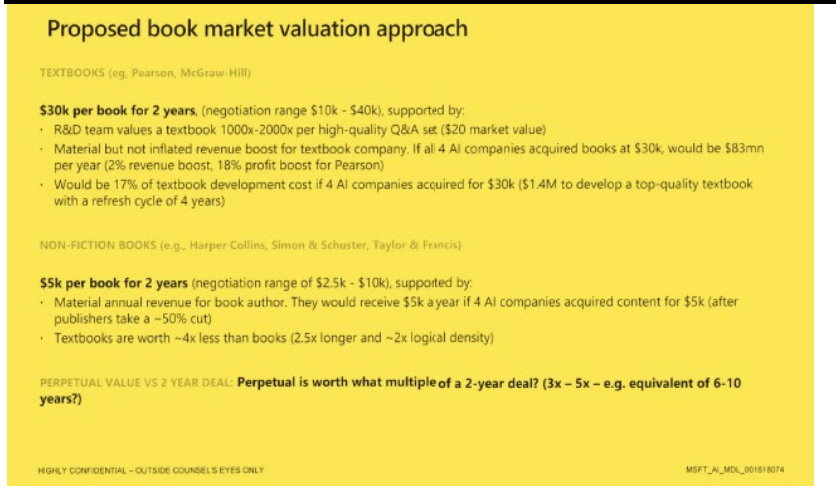
	don't speak for OpenAI." Ex. 671 69 (B. Hecht 30(b)(1) Tr.) at 100:11–12. His statements are not based on personal knowledge of OpenAI, are not attributable to OpenAI, and are inadmissible hearsay as to OpenAI. *See* Fed. R. Civ. P. 56(c)(2), (c)(4); Fed. R. Evid. 602, 801(d)(2)(D), 802. This paragraph cannot be offered to establish any fact concerning OpenAI.
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.	773. **Disputed – Not Supported by Cited Evidence; Not Supported by Admissible Evidence.** Defendants respond that the facts set forth in Paragraph 773 are disputed because they are not supported by the cited evidence. Clarkesworld did not "discontinue[] its fiction competition"; it closed submissions to "give ourselves some breathing room to deal with the real submissions that had come in[.]" Books Ex. 734 at 1–2. Nor were AI-generated submissions the "overwhelming majority," or "many times more" than Clarkesworld could evaluate: its editor reported receiving "700 legitimate submissions and 500 machine-written ones." *Id.* at 2. Defendants also respond that Paragraph 773 is disputed on the grounds that it is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the evidence cited is an unauthenticated transcript of a news interview with a non-party and is inadmissible hearsay. Books Ex. 734 at 3 ("Accuracy and availability of NPR transcripts may vary.").

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.	774. **Disputed – Not Supported by Admissible Evidence.** Defendants respond as follows: The facts set forth in Paragraph 774 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Exhibits 737, 735, 125, and 842 are hearsay and lack foundation because these are non-party publications offered for the truth of their contents, unaccompanied by any authenticating testimony or sponsoring expert to establish their reliability and Plaintiffs identify no exception to the hearsay rule that would apply. These exhibits therefore may not be considered in support of summary judgment. Further, this purported fact is argumentative and not supported by the admissible evidence. [redacted] C. Blythe Opp. Decl. Ex. 70 (C. Sakre 30(b)(6) Tr.) at 58:25-59:10, 73:19-74:10, 177:18-178:1, 182:21-184:1, 222:4-11. Defendants also respond that the facts in Paragraph 774 are immaterial and therefore disputed because these facts do not "affect the outcome of the suit under the governing law." Any alleged harm from so-called "market dilution" from AI-generated books is irrelevant because that

	is not a legally cognizable harm under copyright law. *Kadrey v. Meta Platforms*, 788 F. Supp. 3d 1026, 1052 (N.D. Cal. 2025); *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1032 (N.D. Cal. 2025); *see* ECF 1692 (Microsoft Opening MSJ Br. re Books) at 30-32; ECF 1983 (OpenAI Opening MSJ Br. re Books) at 31-34. The cited support references "*infra* § VII.D" which does not itself constitute evidence; to the extent the facts asserted in those sections are individually disputed, so too is this summary assertion.

B. Defendants' Book-Licensing Agreements

775. [redacted]	775. **Undisputed – Immaterial.** Microsoft does not dispute that Ex. 739 contains the quoted language, but Microsoft responds that the fact stated in Paragraph 775 is immaterial because this fact does not affect the outcome of the suit under the governing law. Whether or not a firm [redacted] has no bearing on any element of the claims or defenses in this case. OpenAI defers to Microsoft's response.
776. [redacted]	776. **Undisputed – Incomplete.** Microsoft does not dispute that Nelson Decl. Exhibit 193 contains the quoted language. But Microsoft disputes Paragraph 776 to the extent it omits the exhibit's full context. Pursuant to Fed. R. Evid. 106, the quoted language appears in a strategy

	team's [redacted] which states that [redacted] Books Ex. 193 at Slide 21. The exhibit further states that [redacted] *Id.* OpenAI defers to Microsoft's response.
777. [redacted]  [redacted]	777. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Exhibit 204 contains the quoted language. But pursuant to Fed. R. Evid. 106, Microsoft also responds that the slide discussed in Paragraph 776 included [redacted] Books Ex. 193, at Slide 21. Microsoft also disputes Paragraph 777 to the extent it implies that [redacted] Books Ex. 204 at MSFT_AI_MDL_001618075. There is "no accepted or reliable mechanism for valuing how one data input affects AI training." ECF 1707 (J. Usdan Decl.) ¶ 11. [redacted]

	[redacted] *Id.* ¶¶ 22, 25. Microsoft further responds that the facts stated in Paragraph 777 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for training generative AI "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
778. [redacted]	778. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Microsoft does not dispute that the Books Ex. 204 at MSFT_AI_MDL_001618074 lists a [redacted] But Microsoft disputes the remainder of the statement because it is not supported by the cited evidence. Microsoft disputes that it [redacted] Books Ex. 204 at MSFT_AI_MDL_001618074 does not discuss licensing content. [redacted] *See, e.g.*, Books Ex. 197 at MSFT_AI_MDL_001286686.

	Microsoft also responds that the facts in Paragraph 778 are immaterial for the reasons stated in its response to Paragraph 777. OpenAI defers to Microsoft's response.
779. [redacted]	779. **Disputed in Part – Immaterial.** Microsoft does not dispute that Exhibit 205 contains the quoted language. But Microsoft disputes that i[redacted] Books Ex. 205, at Slides 13-14 does not discuss licensing content. [redacted] *See, e.g.*, Books Ex. 197 at MSFT_AI_MDL_001286686. Microsoft also disputes the characterization in Paragraph 779 of that language [redacted] Books Ex. 205 at Slides 13–14. Microsoft also responds that the facts set forth in Paragraph 779 are immaterial for the reasons stated in its response to Paragraph 777. OpenAI defers to Microsoft's response.
780. [redacted]	780. **Disputed in Part – Immaterial.** Microsoft does not dispute that Exhibit 209 contains the quoted language. But Microsoft disputes Paragraph 780 to

[redacted]	the extent it suggests that the exhibit [redacted] Books Ex. 49 (G. Russell 30(b)(6) Tr.) 177:14-18; 179:6-8. She further testified that [redacted] *Id.* at 180:20-24; *see also* ECF 1707 (J. Usdan Decl.) ¶ 6. Microsoft also responds that the facts set forth in Paragraph 780 are immaterial for the reasons stated in its response to Paragraph 777. OpenAI defers to Microsoft's response.
781. [redacted]	781. **Disputed – Immaterial.** Microsoft disputes that the September 2023 presentation contemplated [redacted] Exhibit 218 at -123 does not discuss licensing agreements. [redacted] *See, e.g.,* Books Ex. 197 at MSFT_AI_MDL_001286686. Microsoft also responds that the facts stated in Paragraph 781 are immaterial for the reasons stated in its response to Paragraph 777.

	OpenAI defers to Microsoft's response.
782. [redacted].	782. **Undisputed – Immaterial.** Microsoft responds that the facts stated in Paragraph 782 are immaterial because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for training generative AI "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
783. [redacted]	783. **Disputed in Part – Immaterial.** Microsoft does not dispute that Exhibit 186 contains the quoted language. But Microsoft disputes any characterization in Paragraph 783 of the quoted language as a Microsoft approach, position, or recognition. The quoted email [redacted] Books Ex. 186 at MSFT_AI_MDL_001052049. In response, Padma Priya Gaggara, who managed the Bing crawl team, [redacted] *Id.* at MSFT_AI_MDL_001052048–49.

	Microsoft also responds that the facts set forth in Paragraph 783 are immaterial for the reasons stated in its response to Paragraph 782. OpenAI defers to Microsoft's response.
784. [redacted]	784. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Exhibit 214 contains the quoted language. But Microsoft disputes any characterization in Paragraph 784 of the quoted language as a formal position adopted by Microsoft. This exhibit is a working draft that seeks [redacted] Books Ex. 214 at Slides 6, 7. Microsoft also responds that the facts set forth in Paragraph 784 are immaterial for the reasons stated in its response to Paragraph 782. OpenAI defers to Microsoft's response.
785. [redacted]	785. **Undisputed – Immaterial.** Microsoft responds that the facts set forth in Paragraph 785 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for training generative AI "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to

	monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
786. [redacted]	786. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Microsoft does not dispute that Nelson Decl. Exhibit 219 contains the quoted figures. But Microsoft disputes that [redacted] Books Ex. 219 at MSFT_AI_MDL_002853831 does not discuss licensing deals. [redacted]. *See, e.g.*, Books Ex. 197 at MSFT_AI_MDL_001286686. [redacted] ECF 1707 (J. Usdan Decl.) ¶¶ 21–22. Microsoft also responds that the facts set forth in Paragraph 786 are immaterial for the reasons stated in its response to Paragraph 785. OpenAI defers to Microsoft's response.
787. [redacted]	787. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Microsoft does not dispute that Nelson Decl. Exhibit 216 contains the cited figures. But Microsoft disputes Paragraph 787 because [redacted]

[redacted]	[redacted] Books Ex. 216 at MSFT_AI_MDL_002790710, MSFT_AI_MDL_002790712–13. Microsoft also responds that the facts set forth in Paragraph 787 are immaterial for the reasons stated in its response to Paragraph 785. OpenAI defers to Microsoft's response.
788. [redacted]	788. **Undisputed – Immaterial.** Microsoft does not dispute that Books Exhibit 204 contains the quoted language. But Microsoft responds that the facts set forth in Paragraph 788 are immaterial for the reasons stated in its response to Paragraph 785. OpenAI defers to Microsoft's response.
789. [redacted]	789. **Disputed in Part – Immaterial.** Microsoft does not dispute that Exhibit 204 contains the quoted language. But Microsoft disputes the characterization in Paragraph 789 of that language. [redacted] Books Ex. 204 at MSFT_AI_MDL_001618057; *see* ECF 1707 (J. Usdan Decl.) ¶ 11.

	Microsoft also responds that the facts set forth in Paragraph 789 are disputed because they are not supported by the cited evidence. The cited pages in the exhibit do not show that [redacted] Books Ex. 204 at MSFT_AI_MDL_001618023, MSFT_AI_MDL_001618033, MSFT_AI_MDL_001618057. Microsoft also responds that the facts stated in Paragraph 789 are immaterial for the reasons stated in its response to Paragraph 785. OpenAI defers to Microsoft's response.
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.	790. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Exhibit 28 contains the quoted language. But Microsoft disputes that it targeted high-quality long-form text including books for "licensing." In the quoted testimony from Books Ex. 28 (J. Huynh 30(b)(6) Tr. 112:18-115:15), "license" was opposing counsel's characterization of the agreements. [redacted] See, e.g., Books Ex. 197 at MSFT_AI_MDL_001286686. Microsoft also responds that the facts stated in Paragraph 790 are immaterial for the reasons stated in its response to Paragraph 785.

	OpenAI defers to Microsoft's response.
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.	791. **Undisputed – Immaterial.** Microsoft does not dispute that it entered the agreements in Paragraph 791. But Microsoft also responds that the facts set forth in Paragraph 791 are immaterial because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for training generative AI "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
792. [redacted]	792. **Undisputed – Immaterial.** Microsoft responds that the facts set forth in Paragraph 792 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
793. [redacted] .	793. **Undisputed – Immaterial.** Microsoft responds that the facts set forth in Paragraph 793 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
794. [redacted]	794. **Disputed in Part – Immaterial.** Microsoft responds that the facts set forth in Paragraph 794

[redacted]	are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
795. [redacted]	795. **Undisputed – Immaterial.** Microsoft responds that the facts in Paragraph 795 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
796. [redacted]	796. **Undisputed – Immaterial.** Microsoft responds that the facts in Paragraph 796 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
797. [redacted]	797. **Undisputed – Immaterial.** Microsoft responds that the facts set forth in Paragraph 797 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
798. [redacted]	798. **Disputed – Not Supported by Cited Evidence; Immaterial.** Microsoft disputes Paragraph 798 because it is not supported by the cited evidence. The agreement states that: [redacted]

	[redacted] Books Ex. 199 at MSFT_AI_MDL_001286910. Microsoft also responds that the facts in Paragraph 798 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
799. [redacted]	799. **Undisputed – Immaterial.** Microsoft responds that the facts set forth in Paragraph 799 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
800. [redacted]	800. **Disputed – Not Supported by Cited Evidence; Immaterial.** Microsoft disputes that [redacted] because that its not supported by the cited evidence. Section 3(f) of the agreement provides that: [redacted]

	[redacted] Books Ex. 196 at MSFT_AI_MDL_001286673. Microsoft also responds that the facts in Paragraph 800 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
801. [redacted]	801. **Undisputed – Immaterial.** Microsoft does not dispute that Microsoft [redacted]. Microsoft also responds that the facts in Paragraph 801 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
802. [redacted]	802. **Disputed in Part – Incomplete; Immaterial.** Microsoft incorporates its response to Paragraph 801. Microsoft does not dispute that the agreement [redacted] Microsoft disputes that [redacted]

[redacted]	[redacted] Books Ex. 197 at MSFT_AI_MDL_001286692. Regarding the second sentence, pursuant to Fed. R. Evid. 106, the full agreement provides in Section 4(e): [redacted] Books Ex. 197 at MSFT_AI_MDL_001286687. Microsoft disputes any implication that this provision reflects a functioning market to license content for training LLMs. That provision is [redacted] *Id.* Microsoft does not dispute that [redacted] *Id.* at MSFT_AI_MDL_001286693. Microsoft does not dispute that the agreement contains the quoted language in the last sentence of Paragraph 802. Microsoft also responds that the facts in Paragraph 802 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.

803. [redacted]	803. **Undisputed – Immaterial.** Microsoft responds as follows: Microsoft does not dispute the statements in Paragraph 803. Microsoft also responds that the facts in Paragraph 803 are immaterial for the reasons stated in its response to Paragraph 791. OpenAI defers to Microsoft's response.
804. [redacted]	804. **Undisputed.** Microsoft does not dispute that Exhibit 63(a) contains the quoted language. OpenAI defers to Microsoft's response.
805. [redacted]	805. **Undisputed – Incomplete.** Microsoft does not dispute that Exhibit 63(a) contains the quoted language. Pursuant to Fed. R. Evid. 106, Microsoft further responds that [redacted]. ECF 1707 (J. Usdan Decl.) ¶ 22; Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) 77:7–8. Paragraph 805 also omits Mr. Usdan's explanation that [redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) 338:13–22. OpenAI defers to Microsoft's response.

806. [redacted]	806. **Undisputed – Incomplete.** Microsoft does not dispute that Exhibit 63(b) contains the quoted language. But pursuant to Fed. R. Evid. 106, Mr. Usdan also testified that such [redacted] Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) 29:15–30:2; *see also id.* 30:21–31:6. OpenAI defers to Microsoft's response.
807. [redacted]	807. **Undisputed – Incomplete.** Microsoft does not dispute that Mr. Usdan testified that [redacted] But pursuant to Fed. R. Evid. 106, Mr. Usdan also testified that, [redacted] Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) 77:14–78:1. He further testified that [redacted] *Id.* 78:21–25, 80:18–25. OpenAI defers to Microsoft's response.
808. [redacted]	808. **Undisputed – Immaterial.** Microsoft does not dispute that Exhibits 196 and 197 contains the provision

[redacted]	in Paragraph 808. But Microsoft responds that the facts set forth in Paragraph 808 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for training generative AI "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
809. [redacted]	809. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial.** Microsoft does not dispute that [redacted], but Microsoft disputes that [redacted] ECF 1825-2 (M. Lasinski Decl. Ex. B) ¶ 279; Books Ex. 200 at MSFT_AI_MDL_001286934. [redacted]. Books Ex. 200 at MSFT_AI_MDL_001286934. Microsoft also responds that the facts set forth in Paragraph 809 are immaterial for the reasons stated in its response to Paragraph 808. OpenAI defers to Microsoft's response.

810. [redacted]	810. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Exhibit 210 contains the words [redacted] But Microsoft disputes any implication in Paragraph 810 that [redacted] Books Exhibit 210 is a working [redacted], and it states that [redacted] Books Ex. 210 at MSFT_AI_MDL_002301864. Microsoft also disputes that [redacted] *See, e.g.,* Books Ex. 197 at MSFT_AI_MDL_001286686. Microsoft also responds that the facts set forth in Paragraph 810 are immaterial for the reasons stated in its response to Paragraph 808. OpenAI defers to Microsoft's response.
811. [redacted]	811. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Microsoft does not dispute that it entered into [redacted] But Microsoft responds that the facts set forth in Paragraph 811 are disputed to the extent it misstates the cited testimony. [redacted]

	[redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 322:4–12, 322:21–323:4. [redacted]. Pursuant to Fed. R. Evid. 106, Mr. Usdan further testified that [redacted] *Id.* at 322:13–17, 329:3–6. Microsoft also responds that the facts set forth in Paragraph 811 are immaterial for the reasons stated in its response to Paragraph 808. OpenAI defers to Microsoft's response.
812. [redacted]	812. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books Exhibit 63(a) contains testimony from Mr. Usdan regarding his estimate on what Microsoft has spent on books for LLM training since 2024. Pursuant to Fed. R. Evid. 106, Microsoft responds that Mr. Usdan, [redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 336:23–337:12.

	Microsoft also disputes the characterization in Paragraph 812 of [redacted]. *See, e.g.*, Books Ex. 197 at MSFT_AI_MDL_001286686. Pursuant to Fed. R. Evid. 106, Mr. Usdan testified that [redacted] Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) 78:21–23, 84:3–8. He also testified that "there is no market," and that [redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) 325:9–11, 326:4–7. Microsoft also responds that the facts set forth in Paragraph 812 are immaterial for the reasons stated in its response to Paragraph 808. OpenAI defers to Microsoft's response.
813. [redacted]	813. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Books Exhibit 824 contains the quoted language. Pursuant to Fed. R. Evid. 106, Mr. Suleyman testified that he is not an economist and has never "studied the market for licensing books." Books Ex. 824 (M. Suleyman 30(b)(1) Tr.) 154:23–155:4. [redacted]

[redacted]	[redacted] By contrast, Microsoft's 30(b)(6) designee on these deals testified that "there is no market," and that [redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) 325:9–11, 326:4–7; Books Ex. 63(b) (J. Usdan 30(b)(1) Tr.) 84:3–8. Microsoft also responds that the facts set forth in Paragraph 813 are immaterial for the reasons stated in its response to Paragraph 808. OpenAI defers to Microsoft's response.
814. OpenAI also [redacted], OpenAI entered into an agreement [redacted]. Nelson Decl. Ex. 335 at -932.	814. **Disputed in Part – Immaterial.** OpenAI responds as follows: Undisputed that OpenAI entered into an agreement with [redacted]. However, disputed that Plaintiffs' characterization of, and inferences drawn from, that agreement establish that OpenAI [redacted] comparable to Plaintiffs' Asserted Works. [redacted] such as the Asserted Works, [redacted]. Further disputed that the [redacted] establishes the existence or value of comparable transactions for Plaintiffs' works; the agreement [redacted]

	[redacted] *See* G. Leonard Opp. Decl. Ex. A ¶¶ 293–298; G. Leonard Opp. Decl. Ex. B. C. Blythe Opp. ¶ 164. OpenAI further responds that Paragraph 814 is immaterial and therefore disputed because the fact does not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Microsoft defers to OpenAI.

C. Other Licensing Activities

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.	815. **Disputed – Immaterial.** Defendants dispute the fact set forth in Paragraph 815. A functioning market for licensing data, including the Asserted Works, for training LLMs did not exist when the models at issue were trained, does not currently exist, and faces considerable challenges to developing. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 89-145; *See* G. Leonard Opp. Decl. Ex. A ¶¶ 287–291. Defendants further respond that Paragraph 815 is immaterial and therefore disputed because the fact does not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because

	this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
816. The practice of licensing data and content for training computer systems predates large language models. Lasinski Decl. Ex. B at ¶ 49.	816. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that agreements regarding the use of data and content for training computer systems that predate the advent of LLMs exist, but do dispute that Plaintiffs' characterization of the cited historical agreements as establishing a "practice" is supported by the cited evidence. Mr. Lasinski cites only several discrete agreements over a period of 20 years (none of which involve trade books). *See* ECF 1825-2 (M. Lasinski Decl. Ex. B) ¶ 49. Defendants also respond that Paragraph 816 is immaterial and therefore disputed because it does not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.*	817. **Disputed in Part – Immaterial.** Defendants do not dispute that there may have been agreements for the Reuters RCV1 corpus in 2000 (later distributed by NIST in 2004) or the Linguistic Data Consortium's "English Gigaword" dataset in 2003. But Defendants dispute Plaintiffs' characterization of these

	examples as an established historical practice of licensing books or trade-book content (like the Asserted Works) for training computer systems. Neither the Reuters RCV1 corpus nor the English Gigaword dataset involves book content; both are news and newswire corpora. ECF 1825-2 (M. Lasinski Decl. Ex. B) ¶ 49. Defendants also respond that the facts set forth in Paragraph 817 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	818. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that AI developers have agreements with third parties that may involve content But Defendants dispute that such arrangements are a "continued and expanded" "practice" of the activity described in Paragraphs 816-817, and that the cited evidence supports that proposition. Defendants also respond that the facts set forth in Paragraph 818 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop

	is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	819. **Disputed in Part – Incomplete; Immaterial.** Microsoft does not dispute that Mr. Usdan gave the quoted testimony. But pursuant to Fed. R. Evid. 106, Mr. Usdan testified that "there was no market. I mean, there is no market," and that "there's not a market for it." C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 325:9–11, 326:4–7. Microsoft also responds that the facts set forth in Paragraph 819 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
820. [redacted] Lasinski Decl. Ex. B at ¶ 54 & Fig. 2.	820. **Disputed in Part – Immaterial.** Defendants do not dispute that Defendants OpenAI and Microsoft, [redacted], have entered into agreements for content. But Defendants dispute the characterization in Paragraph 820 that such activity is [redacted]

	[redacted]. The existence of recent agreements between AI developers and a few content providers does not imply that there are or will be functioning markets for licensing data for training general-purpose LLMs. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 89-145; G. Leonard Opp. Decl. Ex. A ¶¶ 287–291. Defendants further respond that the facts set forth in Paragraph 820 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	821. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** OpenAI responds as follows: Undisputed that OpenAI has entered into agreements involving access to data for specified AI-related uses. However, disputed to the extent Plaintiffs claim the "content verticals" covered by OpenAI's agreements include trade books like Plaintiffs' Asserted Works. Specifically, the cited evidence identifies agreements OpenAI entered into, but those agreements do not involve access to trade books for LLM training. *See* G. Leonard Opp. Decl. Ex. B ¶¶ 124–27. The disputed proposition is therefore not supported by the record.

	OpenAI further responds that the facts set forth in Paragraph 821 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Microsoft defers to OpenAI.
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.*	822. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Undisputed that OpenAI has entered into agreements involving access to data for specified AI-related uses. However, disputed to the extent Plaintiffs suggest those agreements are comparable to those Plaintiffs contend they could have entered for their Asserted Works. Specifically, the cited evidence does not establish that these agreements include trade books comparable to Plaintiffs' Asserted Works or otherwise establish comparable transactions involving such works. *See* ECF 1825-2 (M. Lasinski Decl. Ex. B) ¶ 55 & Attachments 3.1–3.2. The disputed proposition is therefore not supported by the cited evidence. OpenAI further responds that the facts set forth in Paragraph 822 are immaterial and therefore disputed because these facts do not affect the outcome of the suit

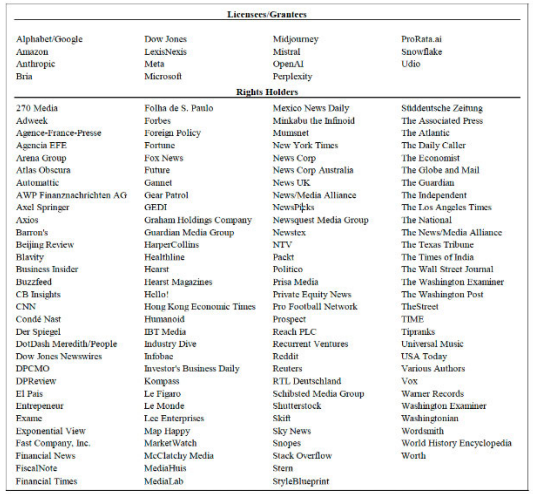
	under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Microsoft defers to OpenAI.
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.	823. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Undisputed that OpenAI entered into the identified agreements. Disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those dollar amounts suggest that consideration was paid solely for access to content for training, the agreements reflect a comparable valuation for Plaintiffs' trade books, or the agreements signal a "market" for Plaintiffs' Asserted works. These discrete agreements include considerations aside from training and did not relate to trade book publishing. *See* ECF 1825-2 (M. Lasinski Decl. Ex. B) Attachments 3.1–3.2; G. Leonard Opp. Decl. Ex. A ¶¶ 287–290. OpenAI further responds that the facts set forth in Paragraph 823 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs

	are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Microsoft defers to OpenAI.
824. [redacted]	824. **Disputed in Part – Immaterial.** Microsoft does not dispute that [redacted] But Microsoft disputes that it [redacted] Books Ex. 197 at MSFT_AI_MDL_001286686; Books Ex. 196 at MSFT_AI_MDL_001286673; *see supra* § VII.B. Microsoft further responds that the facts set forth in Paragraph 824 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. OpenAI defers to Microsoft's response.
825. [redacted]	825. **Disputed in Part – Immaterial.** Defendants do not dispute that [redacted]

[redacted]	[redacted]. However, [redacted] *See* G. Leonard Opp. Decl. Ex. A ¶¶ 342–343; ECF 1726 (C. Tucker Decl.) ¶¶ 124, 128(c). Defendants also respond that the facts set forth in Paragraph 825 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
826. [redacted]	826. **Disputed in Part – Immaterial.** Defendants do not dispute that [redacted] But Defendants respond that the facts set forth in Paragraph 826 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.

827. Beyond the confidential agreements produced in discovery, the AI content licensing market is also reflected in a substantial volume of publicly reported deals.	827. **Disputed – Not Supported by Cited Evidence; Immaterial.** Defendants dispute the facts set forth in Paragraph 827 because Plaintiffs do not cite any evidence to support Paragraph 827. Regardless, Defendants dispute that an "AI content licensing market" exists. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 89-145; G. Leonard Opp. Decl. Ex. A ¶¶ 287–291. Defendants also respond that the facts set forth in Paragraph 827 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	828. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence; Immaterial.** Defendants do not dispute there is public reporting on agreements between certain AI developers and third parties involving access to content for certain AI-related uses. However, disputed to the extent Plaintiffs characterize all of these as "licenses." Defendants also dispute Paragraph 828 to the extent Plaintiffs aggregate agreements involving different types of content, rights, uses, and commercial arrangements as evidence of a single market applicable to Plaintiffs' trade books, as evidence that comparable transactions exist for Plaintiffs' Asserted

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	Monument	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.

Works, or as evidence that Plaintiffs could have entered into comparable transactions for the uses at issue. *See* G. Leonard Opp. Decl. Ex. B ¶¶ 124-126; ECF 1726 (C. Tucker Decl.) ¶¶ 11(e), 89, 128-130. The cited evidence does not establish that comparable transactions exist for Plaintiffs' Asserted Works or that Plaintiffs could have entered into comparable transactions for the uses at issue; rather, the agreements cited by Lasinski involve news publishers, academic publishers, and other content providers, not trade books such as the Asserted Works. *See* G. Leonard Opp. Decl. Ex. B ¶¶ 124-125; ECF 1726 (C. Tucker Decl.) ¶¶ 11(e), 89, 128-130. For these reasons, the disputed proposition is not supported by the cited evidence.

Moreover, Defendants respond that this assertion relies on layers of hearsay (an expert report that purports to summarize agreements based on public reporting). Therefore, this assertion is not supported by admissible evidence.

Defendants further respond that the facts set forth in Paragraph 828 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.

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.	829. **Undisputed – Immaterial.** Undisputed that OpenAI entered into the stated agreements. But OpenAI further responds that the facts set forth in Paragraph 829 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Microsoft defers to OpenAI.
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.	830. **Disputed – Immaterial.** Defendants dispute that recent aggregators and intermediaries "demonstrate[] the development of market infrastructure to facilitate licensing between content owners and AI developers." The existence of aggregators and intermediary platforms does not mean that there is a functioning market for licensing data for training general-purpose LLMs. None of the cited entities has been shown to operate a marketplace or exchange involving trade books (like the Asserted Works) for LLM training, and the existence of start-ups seeking to facilitate transactions involving content does not establish a functioning market in which Plaintiffs could participate. Further disputed to the extent Plaintiffs rely on entities involving different content types, or on investment, funding, or acquisition activity in the purported "ecosystem," because that evidence does not establish

	comparable transactions involving Plaintiffs' trade books. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 131-136; G. Leonard Opp. Decl. Ex. B ¶¶ 136–144. Defendants also respond that the facts set forth in Paragraph 830 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	831. **Disputed in Part – Immaterial.** Defendants do not dispute that the listed entities attempt to facilitate transactions for content for use in generative AI products. But Defendants dispute that these entities "aggregate content for AI licensing." The existence of these entities does not mean that there is a functioning market for licensing data for training general-purpose LLMs. None of the cited entities have been shown to operate a marketplace or exchange involving trade books (like the Asserted Works) for LLM training, and the existence of start-ups seeking to facilitate transactions involving content does not establish a functioning market in which Plaintiffs could participate. Further disputed to the extent Plaintiffs rely on entities involving different content types, or on investment, funding, or acquisition activity in the purported "ecosystem," because that evidence does not establish comparable

	transactions involving Plaintiffs' trade books. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 131-136; G. Leonard Opp. Decl. Ex. B ¶¶ 136–144. Defendants also respond that the facts set forth in Paragraph 831 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
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.	832. **Disputed – Immaterial.** Defendants dispute that these entities have solved challenges to the emergence of a functioning market for licensing trade books for training general purpose LLMs, such as transaction costs. These entities should be treated as experiments rather than evidence of either a solution to the collective action problem or a functioning market for the acquisition of data for AI uses. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 131-136; G. Leonard Opp. Decl. Ex. B ¶¶ 138-142; 212. Additionally, other challenges to a market forming still exist, such as holdout problems and difficulties inherent to pricing or attributing value to any piece of content. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 13, 107-115; G. Leonard Opp. Decl. Ex. B ¶¶ 138-142; 212.

	Defendants also respond that the facts set forth in Paragraph 832 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.
833. Created by Humans launched an AI rights licensing platform for authors in January 2025. Nelson Decl. Ex. 121.	833. **Undisputed – Immaterial; Not Supported by Admissible Evidence.** Defendants do not dispute that Exhibit 121 claims that Created by Humans launched a platform for authors in January 2025 that purports to assist with "AI rights licensing." But the facts set forth in Paragraph 833 are immaterial because they do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Defendants also respond that the facts set forth in Paragraph 833 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the evidence cited is hearsay (a website article).

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.	834. **Undisputed – Immaterial.** Defendants do not dispute that Created by Humans attempts to connect rights holders with AI developers and permits rights holders to grant certain rights. That said, Plaintiffs do not cite any evidence that Created by Humans executed any licensing agreements for training rights. Defendants incorporate their response to Paragraph 833 as to why the existence of Created by Humans is immaterial.
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.	835. **Disputed in Part – Immaterial.** Defendants do not dispute that Created by Humans partnered with authors and entered into certain agreements with publishers and book agencies. However, Defendants dispute Plaintiffs' characterization of, and inferences drawn from, those arrangements to suggest that Created by Humans successfully facilitated transactions between authors or publishers and AI developers for the uses at issue. Singerman testified that [redacted] C. Blythe Opp. Decl. Ex. 72 (J. Singerman 30(b)(6) Tr.) at 114:21–115:12. Defendants also dispute that Created by Humans has [redacted] because Created by Humans [redacted], preventing Defendants from verifying the [redacted]. *Id.* at 50:4-9.

	Defendants also incorporate their response to Paragraph 833 as to why the existence of Created by Humans is immaterial.
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.*	836. **Disputed in Part – Immaterial.** Defendants do not dispute that Created by Humans entered [redacted] However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, those arrangements suggest that Created by Humans successfully facilitated transactions between authors or publishers and AI developers for the uses at issue. Singerman testified that [redacted]. C. Blythe Opp. Decl. Ex. 72 (J. Singerman 30(b)(6) Tr.) at 114:21–115:12. Defendants also dispute that Created by Humans has [redacted] because Created by Humans [redacted], preventing Defendants from verifying the [redacted]. *Id.* at 50:4-9. Defendants also incorporate their response to Paragraph 833 as to why the existence of Created by Humans is immaterial.
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	837. **Disputed in Part – Immaterial.** Defendants do not dispute that Anthropic entered into a class action settlement agreement with plaintiffs in *Bartz* in September 2025. However, disputed to the extent Plaintiffs

books from Library Genesis and the Pirate Library Mirror. Nelson Decl. Ex. 758.	contend the agreement reflects a license rather than a settlement of copyright infringement claims. The cited evidence is a class action settlement agreement, not a content licensing transaction. Books Ex. 758; C. Tucker Opp. Decl. ¶¶ 10-14. Defendants also respond that the facts set forth in Paragraph 837 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Whether a market for licensing data for generative AI purposes "exists or is likely to develop is irrelevant, because this market is not one that plaintiffs are legally entitled to monopolize." *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032. Regardless, a settlement agreement in litigation is not evidence that a market to license content for training LLMs exists. *See, e.g.*, *Beastie Boys v. Monster Energy Co.*, 983 F. Supp. 2d 338, 376 (S.D.N.Y. 2013).
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.	838. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Defendants do not dispute that the Anthropic/Bartz settlement established a $1.5 billion settlement fund and did not grant Anthropic a license for future use. Defendants dispute that the settlement "released Anthropic for past use of the works" because the agreement release states that it covers claims "arising from Defendant's *alleged* past" conduct. Books Ex. 758 § 1.29 (emphasis added). The settlement agreement expressly provides that "[t]he release does not constitute, in

	any respect, a license to torrent, scan, or train AI models on any copyrighted works, or to create infringing outputs from AI models." Books Ex. 758 § 1.29. Defendants also dispute that the settlement fund "corresponds to approximately $3,100 per work" because that is not supported by the cited evidence. The settlement provides for certain deductions from the settlement fund prior to any disbursement of claims, including a "Fee Award" to class counsel that can be up to 25% of the settlement amount ($375,000,000), a service award of $150,000 to the named plaintiffs, and settlement administration expenses. Books Ex. 758 §§ 1.34, 8.1. Thus, any award per book will be less than $3,100 after making those reductions from the settlement fund. Defendants also respond that the facts set forth in Paragraph 838 are immaterial for the reasons stated in Paragraph 837.
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.	839. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that by April 16, 2026, rights holders had filed claims for 440,490 works. For the reasons stated in response to Paragraph 839, Defendants dispute that this equates to $3,400 per claimed work because that is not supported by the cited evidence.

	Defendants also respond that the facts set forth in Paragraph 839 are immaterial for the reasons stated in Paragraph 837.
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.	840. **Disputed in Part – Immaterial.** Defendants do not dispute that as of May 2026, "the claims rate ha[d] reached at least 92.77% of the works." Books Ex. 760 at 7. But Defendants also respond that the facts set forth in Paragraph 840 are immaterial for the reasons stated in Paragraph 837.

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.*	841. **Undisputed – Expert Opinion Subject to Rule 702; Immaterial.** Defendants respond as follows: Undisputed that Amazon is a major retailer of books or that the Lasinski states these facts. But the evidence cited in support of the disputed proposition is currently subject to a motion to exclude expert testimony pursuant to Fed. R. Evid. 702. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Defendants also respond that the facts set forth in Paragraph 841 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Any alleged harm from so-called "market dilution" from AI-generated books is irrelevant because that is not a legally cognizable harm under

	copyright law. *See* ECF 1692 (Microsoft Opening MSJ Br. re Books), at 30-32; ECF 1983 (OpenAI Opening MSJ Br. re Books) at 31-34.
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.	842. **Disputed – Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 842 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is an out-of-court Reuters.com news article offered for the truth of the matters asserted therein and constitutes inadmissible hearsay. *See* Books Ex. 123. Further, even if the cited article were admissible, the assertion is disputed because it does not establish that any books generated using ChatGPT and sold on Amazon are substitutes for the Asserted Works. Book Plaintiffs' own factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." ECF 1983 (OpenAI Opening MSJ Br. re Books) at 35; J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Microsoft defers to OpenAI.
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	843. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Shan produced 13,353 outputs with a median length of approximately 11,266 words. However, disputed that these outputs constitute "book-length output"

books with a median length of approximately 11,266 words. Shan Decl. Ex. D at ¶ 57.	or "multi-chapter books." The disputed proposition is not supported by the cited evidence. Specifically, OpenAI's expert, Dr. Taylor Berg-Kirkpatrick, found that the average word count of Dr. Shan's generated outputs was approximately 12,757 words — far shorter than a typical novel, noting that "for reference, the median word count of the Asserted Works is approximately 119, 728 words." T. Berg-Kirkpatrick Opp. Decl. ¶ 38. Further, OpenAI's expert John Rodzvilla has observed that the word count for books published by Penguin Random House is almost never less than 80,000 words. J. Rodzvilla Opp. Decl. Ex. A ¶ 64. Rodzvilla further notes that certain of the claimed "chapters" of the generated materials contain no more than several lines of text noted as the "summary" for the chapter and an outline briefly noting the sequence of "scenes" that will take place. *Id*. Microsoft defers to OpenAI.
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*.	844. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that the referenced corpus was generated in roughly a few days of automated application programming interface calls. However, disputed that generating each output required no human intervention beyond issuing the initial prompt. The disputed proposition is not supported by the cited evidence. Specifically, Book Plaintiffs' own expert, Dr. Shan, explained that "[e]ach output was produced by chaining four prompt stages: an outline

	prompt, follow-up prompts that pull out individual scenes, chapter-draft prompts, and continuity-tracking prompts that keep characters and plot consistent across chapters." ECF 1814-2 (S. Shan Decl. Ex. B) ¶ 56. Further disputed to the extent Plaintiffs suggest that these outputs constitute "books" that compete with the Asserted Works. J. Rodzvilla Opp. Decl. Ex. A ¶ 63 ("I disagree these outputs are either 'book-length' or 'coherent' in the way that I would expect for books that an average reader would buy to read."). As Book Plaintiffs' factor-four expert admitted, he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." ECF 1983 (OpenAI Opening MSJ Br. re Books) at 35. Microsoft defers to OpenAI.
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.	845. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as to the time and cost spent to generate the referenced material. However, disputed that the output constitutes "a complete multi-chapter book" in any relevant sense. The disputed proposition is not supported by the cited evidence. Specifically, as OpenAI's publishing expert observed, these made-for-litigation outputs "lack coherent plots through which to engage a reader" and "lack developed or cohesive characters." J. Rodzvilla Opp. Decl. Ex. A¶ 63. To the extent that Plaintiffs suggest these outputs from GPT-4o are substitutes for the Asserted Works, Book

	Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." ECF 1983 (OpenAI Opening MSJ Br. Re Books) at 35. This fact is further disputed for the reasons set forth in Defendants' responses to Paragraphs 845–848. Microsoft defers to OpenAI.
846. At that price and speed, producing book-length output at scale is within the reach of an ordinary user. *Id.*	846. **Disputed.** OpenAI responds as follows: This fact is disputed because Book Plaintiffs' technical expert, Dr. Shawn Shan, holds a Ph.D. in computer science, generated the purported "multi-chapter books" described in SUF 845 and used his specialized technical knowledge of API access and prompt engineering to generate such "books." ECF 1814-1 (S. Shan Decl. Ex. A) ¶¶ 1-3, 8; ECF 1814-2 (S. Shan Decl. Ex. B) ¶¶ 56-57; C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) at 190:24-192:4. Specifically, Dr. Shan generated "the entire 14-author corpus" through "automated API calls" over the course of several days. ECF 1814-2 (S. Shan Decl. Ex. B) ¶ 57. An "ordinary user" would not have such abilities. Dr. Shan himself testified that he had not conducted "any studies" of "how often, if at all, everyday users of the models in dispute use prompt sequences to generate multichapter books." C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) at 226:10-22. He further conceded that he had "done, like, no user-specific analysis" in his report, stating only that he believed the techniques were "generally public, easily accessible." *Id.* 123:13-17. Book Plaintiffs have not

	identified any evidence that ChatGPT has been used by any ordinary user to generate book-length output in this manner. Microsoft defers to OpenAI.
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.	847. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 847 are disputed because they are not supported by the cited evidence. To the extent Book Plaintiffs suggest that GPT-4o outputs are substitutes for the Asserted Works, that characterization is wrong. The evidence shows that AI-containing books "garner substantially less usage per title and receive lower star ratings" than human-authored books, and that "direct evidence shows AI-containing books have lower quality, and their rising share … drives the overall decline" in average book quality. Books Ex. 762 at 2–3. Moreover, "an LLM cannot write long-form narratives that automatically balance coherence, quality, thematic consistency, and plot development across thousands of words." Books Ex. 764 at 3. Additionally, Paragraph 847 depends on an arbitrary measure of "quality" set forth by Book Plaintiffs' technical expert that is untethered to any accepted scientific or academic standard; Dr. Shan's methodology used a large language model—Google's Gemini—as an "automated judge of text quality," and his scoring rubric for "writing quality … narrative coherence …, and plot quality" does not correspond to any established benchmark or peer-reviewed

	standard for evaluating book-length creative works. *See* ECF 1814-4 (S. Shan Decl. Ex. D) ¶ 59; *see also* J. Rodzvilla Opp. Decl. Ex. A ¶ 63 n.146 (noting that AI tools are not commonly or reliably used in the industry to evaluate book quality). Microsoft defers to OpenAI.
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.	848. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 848 are disputed because they are not supported by the cited evidence. The assertion that "the best model-generated books matched or exceeded the lowest-scoring published novels by Book Plaintiffs" depends on an arbitrary measure of "quality" conducted by Book Plaintiffs' technical expert using Google's Gemini model as an automated judge. ECF 1814-4 (S. Shan Decl. Ex. D) ¶ 59. That methodology is not a recognized standard for evaluating the quality of book-length creative works; generative artificial intelligence tools "are not commonly used in the industry for the purpose of evaluating the quality of a book," and there is no known "reliable use of the tools for such purpose." J. Rodzvilla Opp. Decl. Ex. A ¶ 63 n.146. Moreover, even the generated materials scoring the highest in Dr. Shan's Gemini test lack a basic plot, do not follow any clear narrative arc or coherent story, and instead present characters trapped in an infinite loop of recurring events with no resolution. J. Rodzvilla Opp. Decl. Ex. A ¶¶ 65–66.

	Microsoft defers to OpenAI.
849. [redacted]	849. **Disputed in Part – Immaterial.** Defendants respond as follows: Undisputed that [redacted] However, disputed to the extent Book Plaintiffs suggest that [redacted] Books Ex. 761 ¶ 5. Defendants also respond that the facts set forth in Paragraph 849 are immaterial because these facts do not affect the outcome of the suit under the governing law. Any alleged harm from so-called "market dilution" from AI-generated books is irrelevant because that is not a legally cognizable harm under copyright law. See ECF 1692 (Microsoft Opening MSJ Br. re Books) at 30-32; ECF 1983 (OpenAI Opening MSJ Br. re Books) at 31-34.

850. [redacted]	850. **Undisputed – Immaterial.** Defendants respond as follows: Undisputed that [redacted] But Defendants also respond that the facts set forth in Paragraph 850 are immaterial for the reasons stated in Paragraph 849 above.
851. [redacted]	851. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: Undisputed as to the figures set forth in fact 851. However, disputed to the extent Book Plaintiffs suggest that [redacted] Further disputed to the extent Book Plaintiffs suggest that [redacted] have caused them harm. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) 99:21-100:1; ECF 1983 (OpenAI Opening MSJ Br. Re Books) at 35.

	Mr. Lasinski testified the same was true with respect to Copilot. J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) 100:18-21. Defendants also respond that the facts set forth in Paragraph 851 are immaterial for the reasons stated in Paragraph 849 above.
852. [redacted]	852. **Disputed – Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 852 are disputed because the cited evidence is incorrect. Disputed that Plaintiffs' assertion that [redacted] is supported by the cited evidence; Lasinski derives that figure by [redacted] Further disputed that [redacted] Books Ex. 761 ¶¶ 5, 6. Accordingly, the cited evidence does not establish that the reported number of AI-generated

	titles understates the amount of AI-generated text on the platform. Defendants also respond that the facts set forth in Paragraph 852 are immaterial for the reasons stated in Paragraph 849 above.
853. [redacted]	853. **Disputed in Part – Immaterial.** Defendants respond as follows: Undisputed that the data produced by [redacted] included these figures for, in Plaintiffs' phrasing, [redacted] Books Ex. 761 ¶¶ 5, 6. Defendants also respond that the facts set forth in Paragraph 853 are immaterial for the reasons stated in Paragraph 849 above.
854. [redacted]	854. **Disputed in Part – Immaterial.** Defendants respond as follows: Undisputed that the data produced by [redacted] included these figures. However, [redacted]

[redacted]	[redacted] Books Ex. 761 ¶¶ 5, 6. Defendants also respond that the facts set forth in Paragraph 854 are immaterial for the reasons stated in Paragraph 849 above.
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.	855. **Disputed in Part.** Defendants respond as follows: Undisputed that Reimers and Waldfogel made the conclusion set forth in Paragraph 855 in their paper, which they concede was not "peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications." Books Ex. 762 at title page. However, disputed to the extent Plaintiffs suggest that ChatGPT or Copilot (as opposed to other AI tools or any other reason) caused any increase in new titles that appear on Amazon. The Reimers and Waldfogel paper makes no causal claim attributing any increase in new titles on Amazon to ChatGPT or Copilot specifically. Further, Reimers and Waldfogel also conclude that books they detected to contain AI text "are less appealing to consumers," "have very little usage," and "are of lower

	quality in the sense of having lower star ratings." *See id.* at 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.*	856. **Disputed in Part.** Defendants respond as follows: Undisputed that Reimers and Waldfogel made the conclusion set forth in Paragraph 856 in their paper, which they concede was not "peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications." Books Ex. 762 at title page. However, disputed to the extent Plaintiffs suggest that ChatGPT or Copilot (as opposed to other AI tools or any other reason) caused any increase of new titles that appear on Amazon. The Reimers and Waldfogel paper makes no causal claim attributing any such increase to ChatGPT or Copilot specifically. Further, Reimers and Waldfogel also conclude that books they detected to contain AI text "are less appealing to consumers," "have very little usage," and "are of lower quality in the sense of having lower star ratings." *See id.* at 3.
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.	857. **Undisputed.**

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.	858. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	859. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	860. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.*	861. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).

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.	862. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	863. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	864. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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*.	865. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
866. The study concluded that "the market now holds many books with substantial AI text that sell little,	866. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF

alongside a smaller set of them that reaches the very top." *Id.* at 3.	2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	867. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	868. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	869. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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"	870. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence,

"where the average drops only because the catalog fills with many lower-revenue AI books." *Id.* at 3–4, 10.	they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	871. **Disputed.** The Court has stricken this 2026 study at Ex. 763 and Lasinski's Supplemental Report relying on such study following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1).
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.	872. **Disputed – Expert Opinion Subject to Rule 702; Immaterial.** Defendants respond as follows: The fact stated in Paragraph 872 is disputed because the evidence cited in support is currently subject to a motion to exclude expert testimony pursuant to Fed. R. Evid. 702. Specifically, Defendants have moved to exclude Lasinski's Rebuttal Report Sections 9.0 through 9.6 (ECF 1825-2 (M. Lasinski Decl. Ex. B) ¶¶ 362–412), which encompass the cited paragraph 366, on the grounds that Lasinski's market-harm opinions lack a reliable methodology and are not supported by sufficient facts or data. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Further, the facts set forth in Paragraph 872 are disputed by Defendants because they are not supported by the cited evidence. The cited study itself concludes that "an LLM cannot write long-form narratives that automatically balance coherence, quality, thematic consistency, and plot development across thousands of words." Books Ex. 764 at 3.

	Defendants also respond that the facts set forth in Paragraph 872 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Any alleged harm from so-called "market dilution" from AI-generated books is irrelevant because that is not a legally cognizable harm under copyright law. *See* ECF 1692 (Microsoft Opening MSJ Br. re Books) at 30-32; ECF 1983 (OpenAI Opening MSJ Br. re Books) at 31-34.
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.	873. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 873 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is inadmissible hearsay. Book Plaintiffs [redacted] *See* C. Blythe Opp. Decl. Ex. 70 [redacted]. Further, Defendants dispute the fact in Paragraph 873 because it does not explain what is meant by "market effects." Defendants also respond that the facts set forth in Paragraph 873 are immaterial for the reasons stated in response to Paragraph 872.
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	874. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 874 are disputed because they are not supported by admissible

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.	evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Defendants dispute that Plaintiffs may rely on the NPR article or Ms. Rasenberger's statements quoted in that article to establish that AI-generated books were crowding Amazon's marketplace, stealing sales, or generating revenue before Amazon identified them. The cited NPR article and the statements it quotes are hearsay offered for their truth. Defendants also respond that the facts set forth in Paragraph 874 are immaterial for the reasons stated in response to Paragraph 872.
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.	875. **Disputed – Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 875 are disputed because they are not supported by the cited evidence. Ms. Rasenberger lacked personal knowledge of this claim. As she confirmed during her deposition, [redacted] C. Blythe Opp. Decl. Ex. 1 (M. Rasenberger 30(b)(6) Tr.) at 22:17–22. Defendants also respond that the facts set forth in Paragraph 875 are immaterial for the reasons stated in response to Paragraph 872.
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.	876. **Undisputed – Immaterial.** Defendants respond as follows: Undisputed that Ms. Rasenberger so testified. Defendants also respond that the facts set forth in

	Paragraph 876 are immaterial for the reasons stated in response to Paragraph 872.
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.	877. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: Paragraph 877 is disputed because Book Plaintiffs offer no support whatsoever for this assertion. Indeed, Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id.* 110:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) at 298:7-299:8. Additionally, Book Plaintiffs have not identified a single book generated by ChatGPT or Copilot that infringes the Asserted Works and is available for sale. ECF 1726 (C. Tucker Decl.) ¶¶ 45-46, 59-60, 87-88. Defendants also respond that the facts set forth in Paragraph 877 are immaterial for the reasons stated in response to Paragraph 872.
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	878. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: Undisputed that Ms. Rasenberger so testified but Defendants dispute that the substance of this

take some sales from the real author's book." *Id.* at 86:21–88:24.	testimony is evidence that AI-generated books have taken sales from human-authored works. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id.* 110:18-21. Furthermore, the disputed proposition is not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, Ms. Rasenberger's testimony is entirely based on hearsay about what agents have purportedly told the Authors Guild. Defendants also respond that the facts set forth in Paragraph 878 are immaterial for the reasons stated in response to Paragraph 872.
879. The Authors Guild was notified by Stephanie Land of an AI-generated mimic of her memoir. *Id.* at 77:2–15.	879. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 879 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Ms. Rasenberger's testimony is entirely based on layers of hearsay. Ms. Rasenberger testified that Stephanie Land [redacted] Books Ex. 46(a) (M. Rasenberger 30(b)(1) Tr.)

	77:5–8; *see also id.* at 79:25–82:1. Defendants also respond that the facts set forth in Paragraph 879 are immaterial for the reasons stated in response to Paragraph 872.
880. Jane Friedman complained of AI-generated books published under her name. Nelson Decl. Ex. 138.	880. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 880 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is inadmissible hearsay, as Exhibit 138 is an email that purports to summarize Jane Friedman's out-of-court complaint regarding AI-generated books. Defendants also respond that the facts set forth in Paragraph 880 are immaterial for the reasons stated in response to Paragraph 872.
881. AI knockoffs of Seth Harp's work appeared on Amazon, including on its Prime Deals savings page. Nelson Decl. Ex. 140.	881. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 881 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the evidence cited is inadmissible hearsay, as Nelson Declaration Exhibit 140 is an email that purports to summarize an X thread about Seth Harp's work appearing on Amazon. Defendants also respond that the facts set forth in Paragraph 881 are immaterial for the reasons stated in response to Paragraph 872.

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.	882. **Disputed – Not Supported by Cited Evidence; Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: The facts set forth in Paragraph 882 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the facts rest on layers of hearsay, including an email that purports to summarize a social media post about purportedly AI-generated works. Defendants also respond that the facts set forth in Paragraph 882 are immaterial for the reasons stated in response to Paragraph 872.
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.	883. **Disputed – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: The facts set forth in Paragraph 883 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the assertion entirely relies on hearsay (newspaper articles that purport to summarize the business plans of Spines). Defendants also respond that the facts set forth in Paragraph 883 are immaterial for the reasons stated in response to Paragraph 872.
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	884. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: Undisputed that the cited evidence contains the statements reported in Paragraph 884. However, Paragraph 884 is disputed as it is not supported by admissible evidence, as required by Local Rule 56.1, and

York Times published a quiz inviting readers to distinguish AI writing from human writing. Nelson Decl. Ex. 120.	therefore is not properly before the Court. Specifically Books Ex. 102, Ex. 115, and Ex. 120 are hearsay. Defendants also respond that the facts set forth in Paragraph 884 are immaterial for the reasons stated in response to Paragraph 872.
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.	885. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: Undisputed that Hachette withdrew the novel Shy Girl after suspected AI use. However, disputed because the March 2026 statement from The New York Times is hearsay, is not admissible evidence, and is therefore not properly before the Court. Defendants also respond that the facts set forth in Paragraph 885 are immaterial for the reasons stated in response to Paragraph 872.
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.	886. **Disputed in Part – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: Undisputed that The New York Times reported on such a story. However, disputed to the extent Book Plaintiffs suggest that this story was generated by ChatGPT or Copilot because the cited exhibit does not support such an assertion. Further disputed because the statement from The New York Times is hearsay, is not admissible evidence, and is therefore not properly before the Court. Defendants also respond that the facts set forth in Paragraph 886 are immaterial for the reasons stated in response to Paragraph 872.

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.	887. **Disputed – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: The facts set forth in Paragraph 887 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, this fact is disputed because this assertion relies on hearsay. Defendants also respond that the facts set forth in Paragraph 887 are immaterial for the reasons stated in response to Paragraph 872.
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.	888. **Disputed – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: The facts set forth in Paragraph 888 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, the article relies on a 2026 study at Ex. 763 that the Court has stricken following Defendants' Motion to Strike. ECF 2090. Because Plaintiffs may not rely on the cited evidence, they cannot support this asserted fact on summary judgment. *See* Fed. R. Civ. P. 56(c)(1). Further disputed because the author of the article stated: "The suggestion that I used generative AI to write Daggermouth is wholly untrue." Books Ex. 125 at 3. Defendants also respond that the facts set forth in Paragraph 888 are immaterial for the reasons stated in response to Paragraph 872.
889. The New York Times reported that author Jess Michaels said "[t]he glut of the A.I. work makes	889. **Disputed – Not Supported by Admissible Evidence; Immaterial**. Defendants

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.	respond as follows: The facts set forth in Paragraph 889 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, this assertion relies on at least two layers of hearsay (a newspaper article that purports to quote an author). Defendants also respond that the facts set forth in Paragraph 889 are immaterial for the reasons stated in response to Paragraph 872.
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.	890. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 890 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, this assertion relies on at least two layers of hearsay (a Wall Street Journal article that purports to summarize or quote what an author's agency wrote in a letter). Defendants also respond that the facts set forth in Paragraph 890 are immaterial for the reasons stated in response to Paragraph 872.
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.	891. **Disputed in Part.** Defendants respond as follows: Undisputed that Branch offered the quoted testimony. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Branch's testimony establish that Book Plaintiffs have suffered actual or potential market harm from Defendants' products. When asked, "Is the market for you currently being displaced by generative AI," Branch responded: "I don't know." C. Blythe Opp. Decl. Ex. 74 (T. Branch 30(b)(1) Tr.) 126:20–

	127:11. Branch further testified that the only evidence he had of lost revenue was "that the books weren't paid for to train." *Id.* at 73:11–74:1. Furthermore, Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id.* 100:18-21. Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16-60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
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.	892. **Disputed in Part.** Defendants respond as follows: Undisputed that Ex. 54 contains the quoted testimony. However, disputed that Plaintiffs' characterization of, and inferences drawn from, Shapiro's testimony establish that Shapiro has suffered actual or potential market harm from Defendants' models. Shapiro's testimony explains that his "significant harm" refers to the *possibility* that students in Shakespeare courses could choose to use summaries of his books instead of purchasing the books he would *hypothetically* recommend. Books Ex. 54 (J. Shapiro 30(b)(1) Tr.) 90:1–91:5. Nor does this testimony establish that any ChatGPT or Copilot output is a substitute for the Asserted Works. Indeed, Book Plaintiffs' factor-four expert admitted that he did not have the "data

	necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id*. 100:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) 298:7-299:8. And Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
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.	893. **Disputed in Part.** Defendants respond as follows: Undisputed that Baldacci testified concerning materials released contemporaneously with his novel *Nash Falls* and an Amazon ranking associated with at least one such item for purchase. However, disputed to the extent Plaintiffs suggest that such materials were purchased as an alternative to, or substitute for, *Nash Falls* or any of Baldacci's Asserted Works. Baldacci testified, [redacted] Books Ex. 5 (D. Baldacci Tr.) 228:2–4. Furthermore, Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale."

	J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id*. 110:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) 298:7-299:8. And Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
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.	894. **Disputed in Part.** Defendants respond as follows: Undisputed that Connelly provided the quoted testimony. However, disputed to the extent Book Plaintiffs suggest that any such outlines have caused them any harm. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id*. 100:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) 298:7-299:8. And Defendants' experts found that there have not been measurable declines

	in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
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.	895. **Disputed in Part.** Defendants respond as follows: Undisputed that Grisham provided the quoted testimony. However, disputed to the extent Book Plaintiffs suggest that any "AI-generated books" have caused them any harm. Grisham [redacted]. Books Ex. 24 (J. Grisham Tr.) 176:22–25. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr. 99:21–100:17). Mr. Lasinski testified the same was true with respect to Copilot. *Id.* 100:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) 298:7-299:8. And Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
896. Stacy Schiff testified that she believes ChatGPT is an existing threat to her livelihood because it	896. **Disputed in Part.** Defendants respond as follows: Undisputed that Schiff provided the quoted

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.	testimony. However, disputed to the extent Book Plaintiffs suggest that any "AI-generated books" have caused them any harm. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the same was true with respect to Copilot. *Id*. 100:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) 298:7-299:8. And Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.
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.	897. **Disputed in Part.** Defendants respond as follows: Undisputed that Franzen testified about [redacted] Books Ex. 17 (J. Franzen 30(b)(1) Tr.) 107:19–108:4. However, disputed that the cited evidence shows that any outputs have harmed Franzen. Book Plaintiffs' factor-four expert admitted that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz. Decl. Ex. 116 (M. Lasinski Tr.) at 99:21–100:17. Mr. Lasinski testified the

	same was true with respect to Copilot. *Id.* 100:18-21. Further, Mr. Seeley, Book Plaintiffs' expert opining on the publishing industry, described potential competition with AI-generated books as, at most, a future harm, admitting that "these harms haven't happened yet[.]" C. Blythe Opp. Decl. Ex. 75 (M. Seeley Tr.) at 298:7-299:8. And Defendants' experts found that there have not been measurable declines in the trends of Plaintiffs' royalties and sales of the Asserted Works following the introduction of ChatGPT and Copilot. *See* ECF 1726 (C. Tucker Decl.) ¶¶ 11, 16–60; ECF 1729 (G. Leonard Decl.) ¶¶ 10-34.

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.	898. **Disputed in Part – Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702; Immaterial.** Defendants respond as follows: Undisputed that the Zhao declaration contains the stated opinions, which are expert opinions and not facts. However, disputed that Dr. Zhao is qualified to opine on OpenAI and Microsoft's knowledge or state of mind, or that his opinions establish OpenAI and Microsoft's knowledge or intent. The evidence cited in support of the disputed proposition is currently subject to a motion to exclude expert testimony pursuant to Fed. R. Evid. 702. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). The evidence cited is expert opinion testimony that does not establish the underlying facts asserted by Plaintiffs.

	Further, Dr. Zhao's opinions concerning general purported risks to labor markets, mental health, education, cybersecurity, and public safety are unreliable and do not establish the asserted facts concerning OpenAI or the models at issue. *See* ECF 1787-2 (B. Zhao Decl. Ex. B) ¶¶ 112–13. Defendants also respond that the statement in Paragraph 898 that "OpenAI and Microsoft were aware of those risks and released their models anyway" is disputed because the cited evidence does not support the claim. The cited paragraphs of Dr. Zhao's report do not state that Defendants "were aware of those risks and released their models anyway." ECF 1787-2 (B. Zhao Decl. Ex. B) ¶¶ 112–13. Defendants further respond that the facts set forth in Paragraph 898 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Non-copyright cognizable harms have no bearing on any claim or defense in this litigation.
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.	899. **Disputed – Not Supported by Cited Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 899 are disputed because they are not supported by the cited evidence. Specifically, the cited evidence is a conditional statement that does not reflect the current state of the world. Plaintiffs' statement that "AI is upending domestic and international labor markets" is an opinion and not a statement of fact. Moreover, Plaintiffs mischaracterize the *International AI*

	Safety Report 2026 to the extent it states with certainty that AI systems with more autonomy "would likely accelerate labour market disruption." Ex. 765 states that there may be "labour market disruption" only "if AI agents gained the capacity to act with greater autonomy across domains" and "reliably manag[ed] longer, more complex sequences of tasks in pursuit of higher-level goals." Books Ex. 765 at 87. Further, the same report states that "[e]arly evidence shows no effect on overall employment," and that new research "found no relationship between an occupation's AI exposure or AI adoption and overall employment." *Id.* at 12, 84. Defendants further respond that the facts set forth in Paragraph 899 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Defendants also respond that the facts set forth in Paragraph 899 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Books Ex. 765 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment.
900. OpenAI's internal documentation acknowledges that artificial general intelligence "could	900. **Undisputed – Immaterial.** OpenAI further responds that the facts set forth in Paragraph 900

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.	are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft defers to OpenAI.
901. Sam Altman testified that "some jobs will go away entirely." Nelson Decl. Ex. 2(b) at 77 :16-78:21.	901. **Disputed in Part – Immaterial.** OpenAI responds as follows: Undisputed that Altman testified that "some jobs will go away entirely." Disputed that Plaintiffs' selective quotation accurately reflects the context of Altman's testimony because pursuant to Fed. R. Evid. 106, it omits his qualification that this would occur "like [what] happens with every other technological revolution." Books Ex. 2(b) (S. Altman 30(b)(6) Tr.) 78:18-21. Microsoft defers to OpenAI.
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.	902. **Undisputed – Immaterial.** Microsoft does not dispute that Books Ex. 21 contains the quoted language. But Microsoft responds that the facts set forth in Paragraph 902 are immaterial and therefore disputed because these facts do not affect the outcome of the suit under the governing law. Non-copyright cognizable harms have no bearing on any claim or defense in this litigation. OpenAI defers to Microsoft's response.
903. LLMs expose vulnerable users to unmitigated harm to their mental health, and AI chatbots "respond inappropriately to various mental health conditions,	903. **Disputed in Part – Not Supported by Cited Evidence; Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows:

encouraging delusions and failing to recognize crises." Nelson Decl. Ex. 767 at 11 ("Conclusion").	Undisputed that Books Ex. 767 contains the quoted statement. However, disputed that Books Ex. 767 supports Plaintiffs' broader assertion that "LLMs expose vulnerable users to unmitigated harm to their mental health." The disputed proposition is not supported by the cited evidence. Specifically, Books Ex. 767 is a single study "investigat[ing] the use of LLMs to *replace* mental health providers," not a study on the general impacts of LLMs on mental health. *See* Books Ex. 767 at 1. Further disputed that Plaintiffs' characterization of, and inferences drawn from, the quoted statement apply to all AI chatbots or LLMs; the cited exhibit addresses "commercially-available therapy bots," not AI chatbots or LLMs generally. *See* Books Ex. 767 at 8. Defendants also respond that the facts set forth in Paragraph 903 are disputed on the grounds that they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. S.D.N.Y. Specifically, Books Ex. 767 is hearsay and lacks foundation because it is a non-party publication offered for the truth of its contents, unaccompanied by any authenticating testimony or sponsoring expert to establish its reliability and Plaintiffs identify no exception to the hearsay rule that would apply. It therefore may not be considered in support of summary judgment. Defendants further respond that the facts set forth in Paragraph 903 are immaterial and therefore disputed for the reasons stated in Paragraph 898.

904. [redacted] at -363.	904. **Disputed in Part – Immaterial; Not Supported by Admissible Evidence.** Microsoft does not dispute that Books Ex. 768 contains the quoted language. But Microsoft disputes the mischaracterizations in Paragraph 904 of Exhibit 768. The document [redacted] Books Ex. 768 at MSFT_AI_MDL_000015355. The Deployment Safety Board was established to conduct "a comprehensive pre-release safety review process" for advanced OpenAI models. ECF 1710 (S. Bird Decl.) ¶ 13. Identifying risk categories for that review is how Microsoft "map[s]" potential risks so they can be measured and mitigated before deployment. *Id.* ¶ 23. Microsoft also disputes Plaintiffs' characterization of, and inferences drawn from, Books Ex. 768 establish that [redacted] Books Ex. 768 at MSFT_AI_MDL_000015363-64. Microsoft also responds that the facts set forth in Paragraph 904 are immaterial and therefore disputed for the reasons stated in Paragraph 898. OpenAI defers to Microsoft's response.

905. [redacted] at -308.	905. **Disputed in Part – Immaterial.** Microsoft does not dispute that Books Ex. 774 contains the quoted language. But Microsoft disputes any implication in Paragraph 905 that the language quoted from the exhibit describes the behavior of a deployed model. The cited exhibit [redacted] Books Ex. 774 at MSFT_AI_MDL_001093308–09. [redacted] *Id.* at MSFT_AI_MDL_001093310. Such "red teaming or stress testing" is "a key mechanism used to intentionally elicit undesirable model behaviors" in order to understand "the limitations and vulnerabilities of the system" before deployment. ECF 1710 (S. Bird Decl.) ¶ 23(a); *see also id.* ¶¶ 12–13, *27*. Microsoft further disputes this paragraph to the extent Plaintiffs' characterization of, and inferences drawn from, Books Ex. 774 suggest that the model spontaneously generated or generally exhibited the identified behavior. The cited statement [redacted] Books Ex. 774 at -308. Further disputed to the extent Plaintiffs suggest that this test result establishes that the model encouraged self-harm outside the particular prompt described in Ex. 774.

	Microsoft also responds that the facts set forth in Paragraph 905 are immaterial for the reasons stated in its response to Paragraph 898. OpenAI defers to Microsoft's response.
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.	906. **Disputed – Not Supported by Admissible Evidence; Immaterial.** OpenAI responds as follows: The facts set forth in Paragraph 906 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Plaintiffs cite allegations in unadjudicated complaints, not admissible evidence establishing the truth of those allegations. OpenAI objects to Plaintiffs' reliance on the allegations for their truth because they constitute inadmissible hearsay. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Further disputed that the allegations are relevant to the claims or defenses at issue, and OpenAI objects under Fed. R. Evid. 401 and 403 to the extent Plaintiffs seek to use allegations from unrelated pending litigation as evidence in this action. OpenAI further responds that the facts set forth in Paragraph 906 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft defers to OpenAI.
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	907. **Disputed – Not Supported by Admissible Evidence; Immaterial**. Defendants respond as follows: The facts set forth in Paragraph 907 are

students reporting that LLMs have replaced the learning process in numerous classes. Nelson Decl. Ex. 771.	disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Defendants object to Plaintiffs' reliance on Books Ex. 771 because the cited material contains inadmissible hearsay or other statements that cannot be presented in admissible form at trial. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Further disputed that these assertions concerning student cheating and educational practices are relevant to the claims or defenses at issue, and Defendants object under Fed. R. Evid. 401 and 403. Further disputed that Books Ex. 771 supports Plaintiffs' assertion that AI-assisted cheating is widespread in both K–12 and higher education; the cited evidence addresses higher education, not K–12 education. *See* Books Ex. 771. Further disputed that the cited evidence establishes that LLMs have "replaced the learning process in numerous classes" for college students, rather than reporting particular students' perceptions or experiences concerning LLM use. *See id*. Defendants further respond that the facts set forth in Paragraph 907 are immaterial and therefore disputed for the reasons stated in Paragraph 898.
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.	908. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: The facts set forth in Paragraph 908 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Defendants object to Plaintiffs' reliance on Books Ex. 771 because the

	cited material, a New York Magazine article, is itself inadmissible hearsay and cannot be presented in admissible form at trial. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Additionally, Plaintiffs' assertions concerning the reliability of AI-plagiarism detection and schools' disciplinary practices are not relevant to any claim or defense in this action, and Defendants object to Plaintiffs' reliance on this evidence, which should be disregarded. Fed. R. Evid. 401; Fed. R. Evid. 403. Defendants further respond that the facts set forth in Paragraph 908 are immaterial and therefore disputed for the reasons stated in Paragraph 898.
909. LLMs are easily weaponized to autonomously plan and execute complex, multi-step network attacks without human instruction. Nelson Decl. Ex. 773.	909. **Disputed – Not Supported by Admissible Evidence; Immaterial.** Defendants respond as follows: Disputed that Plaintiffs' characterization that LLMs are "easily weaponized to autonomously plan and execute complex, multi-step network attacks without human instruction" is supported by the cited evidence. This paragraph is disputed to the extent Plaintiffs' characterization of, and inferences drawn from, Books Ex. 773 improperly generalize results from specific cybersecurity evaluations to LLMs generally. The cited evidence describes experimental results of a research prototype which the researchers admit "only works under specific conditions." *See* Books Ex. 773 at 4. Further, the facts set forth in Paragraph 909 are not supported by admissible evidence, as required by Local Rule 56.1, and therefore are not properly before the Court. Specifically, Defendants object to Plaintiffs' reliance on Books Ex. 773

	because the cited material, a Carnegie Mellon College of Engineering article, is itself inadmissible hearsay and cannot be presented in admissible form at trial. Fed. R. Evid. 802; Fed. R. Civ. P. 56(c)(2). Additionally, Defendants object to Plaintiffs' reliance on this evidence under Fed. R. Evid. 401 and 403 because assertions concerning hypothetical cybersecurity risks are not relevant to the claims or defenses at issue. Defendants further respond that the facts set forth in Paragraph 910 are immaterial and therefore disputed for the reasons stated in Paragraph 898.
910. [redacted] at -316.	910. **Disputed in Part – Immaterial.** Microsoft does not dispute that Exhibit 774 contains the quoted language. But Microsoft disputes any implication in Paragraph 910 that the exhibit language quoted describes a deployed model. [redacted] Ex. 774 at MSFT_AI_MDL_001093308–09. [redacted]. *Id.* at MSFT_AI_MDL_001093310. [redacted] *Id.* at MSFT_AI_MDL_001093316. Such "red teaming or stress testing" is used "to intentionally elicit undesirable model behaviors" before deployment. ECF 1710 (S. Bird Decl.) ¶ 23(a). Further disputed to the extent Plaintiffs suggest that the identified evaluation establishes actual harm or conduct outside the testing described in Ex. 774.

	Microsoft also responds that the facts set forth in Paragraph 910 are immaterial for the reasons stated in Paragraph 898. OpenAI defers to Microsoft's response.
911. [redacted]	911. **Disputed in Part – Immaterial; Not Supported by Admissible Evidence.** Microsoft does not dispute that Exhibit 774 states that "DV3 can be prompted" as stated. *See* Books Ex. 774 at MSFT_AI_MDL_001093313. The document [redacted] *Id.* [redacted] *Id.* at MSFT_AI_MDL_001093314. [redacted]. *Id.* at MSFT_AI_MDL_001093308–10. Such "red teaming or stress testing" is used "to intentionally elicit undesirable model behaviors" before deployment. ECF 1710 (S. Bird Decl.) ¶ 23(a). Thus, the cited statements do not establish that [redacted]. *See* Books Ex. 774 at MSFT_AI_MDL_001093313-14. Further disputed to the extent Plaintiffs suggest that Microsoft's identification of a potential danger establishes actual harm or misuse outside the evaluation context.

	Microsoft also responds that the facts set forth in Paragraph 911 are immaterial for the reasons stated in its response to Paragraph 898. OpenAI defers to Microsoft's response.
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.	912. **Disputed in Part – Immaterial**. OpenAI responds as follows: Undisputed that Ex. 775 contains the quoted statement. However, disputed to the extent Plaintiffs suggest that OpenAI's LLMs have been used in this manner. The cited evidence refers to hypothetical risks, not actual events. Books Ex. 775 at OPCO_MDL_001666700–701. The document states that "[p]art of the reason [OpenAI] exist[s] is to prevent" the misuse of AI tools. Books Ex. 775 at OPCO_MDL_001666699. OpenAI further objects to Plaintiffs' reliance on this evidence under Fed. R. Evid. 401 and 403 because assertions concerning hypothetical misuses of AI are not relevant to the claims or defenses at issue. OpenAI further responds that the facts set forth in Paragraph 912 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft defers to OpenAI.
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.	913. **Disputed in Part – Immaterial**. OpenAI responds as follows: Undisputed that Ex. 776 contains the quoted statement. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that statement suggest any such action has taken place. The

	cited evidence refers to hypothetical risks, not actual events. *See* Books Ex. 776 at OPCO_MDL_004775334. Further disputed to the extent Plaintiffs omit the document's discussion of measures to address such hypothetical risks, including that "we can improve our security and ramp up confidentiality measures, we can predict misuse and prepare mitigations, [and] we can improve our robustness to jailbreaks." OpenAI objects to Plaintiffs' reliance on this evidence under Fed. R. Evid. 401 and 403 because assertions concerning hypothetical misuses of AI are not relevant to the claims or defenses at issue. OpenAI further responds that the facts set forth in Paragraph 913 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft defers to OpenAI.
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.	914. **Disputed in Part – Not Supported by Cited Evidence; Immaterial.** Defendants do not dispute that Nelson Decl. Exhibit 752 contains the quoted language. But Defendants respond that the first statement ("Defendants' leadership acknowledges that continued AI development poses societal-scale risks, and Defendants released their models anyway") is not supported by the cited evidence. The statement on AI risk does not say that "continued AI development poses societal-scale risks," does not attribute any risk to OpenAI's models, and says nothing about the release of any model. *See* Books Ex. 752. Defendants also dispute that Plaintiffs' characterization of,

	and inferences drawn from, the statement establish that "Defendants' leadership acknowledges that continued AI development poses societal-scale risks." The cited evidence demonstrates that OpenAI and Microsoft personnel agreed that risks should be mitigated. *See* Books Ex. 752. Further disputed that Plaintiffs' characterization that Defendants "released their models anyway" is supported by Ex. 752, which addresses mitigation of potential AI risks rather than whether AI models should not be developed or released. Defendants also respond that the facts set forth in Paragraph 914 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft also responds as follows: Microsoft disputes any implication in Paragraph 914 that Bill Gates was part of Microsoft leadership acting on Microsoft's behalf and/or with decision-making authority over Microsoft's decisions. *See* Response to Paragraph 187.
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.	915. **Disputed in Part; Immaterial.** OpenAI responds as follows: Undisputed that Ex. 766 contains the quoted text. However, disputed to the extent Plaintiffs' characterization of, and inferences drawn from, that statement suggest that the development of artificial general intelligence necessarily will result in "autonomous or difficult-to-control actions" or a threat to humanity. The cited language identifies a potential risk associated with future AGI, not an actual characteristic of existing models.

	See Books Ex. 766 at OPCO_MDL_001008367. Further disputed to the extent Plaintiffs omit the document's discussion of measures intended to mitigate potential misuse of AI tools. *See* Books Ex. 766 at OPCO_MDL_001008380-85. OpenAI objects to Plaintiffs' reliance on this evidence under Fed. R. Evid. 401 and 403 because assertions concerning hypothetical misuses of AI are not relevant to the claims or defenses at issue. OpenAI further responds that the facts set forth in Paragraph 915 are immaterial and therefore disputed for the reasons stated in Paragraph 898. Microsoft defers to OpenAI.
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.	916. **Undisputed – Immaterial.** Microsoft defers to OpenAI.

VIII. Copies of Plaintiffs' Works Are Present in OpenAI's Data[1]

917. Plaintiffs' Expert, Dr. Shawn Shan, identified Plaintiffs' Works in Defendants' data using two primary methods. Shan Decl. Ex. A ¶¶ 23–36.	917. **Disputed in Part – Immaterial; Not Supported by Admissible Evidence; Improper Legal Argument.** Defendants do not dispute that Dr. Shan used multiple methodologies to purportedly identify Plaintiffs' Works or portions of those works in "Defendants' Data," but Defendants dispute that this vague description of

[1] For purposes of summary judgment, OpenAI does not contest the elements of *prima facie* infringement as to the 194 Asserted Works that Book Plaintiffs have sought summary judgment on. OpenAI's alleged use of these works is a fair use as a matter of law. OpenAI continues to reserve all defenses, including with respect to other works not at issue in Book Plaintiffs' motion.

	Shan's methodology is a material fact as it does not affect the outcome of the suit under the governing law. Further, Defendants dispute this fact's characterization of "Defendants' data" as it treats various distinct, unrelated data sets from two distinct parties as one set of "data." Defendants further dispute that those methods "identified Plaintiffs' Works in Defendants' data" as one of the two methods involved an analysis of metadata, which does not represent the actual text of any of the Asserted Works. Nothing in the cited paragraphs explains how metadata was used by Dr. Shan to "identify" any of the Asserted Works in Defendants' data. Also disputed as stating a legal conclusion to the extent this statement suggests that Defendants' data contained unauthorized or unlawful copies or reproductions of any Asserted Works under the meaning of copyright law. As the following statements explain, the "hash-matching" method used by Dr. Shan only searched for portions of an Asserted Work. Further, the Court should disregard the disputed portion of Paragraph 917 because the assertion constitutes improper attorney argument and is conclusory. It thus requires no response.
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.	918. **Disputed in Part – Immaterial; Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that the cited paragraphs from Dr. Shan's Opening Report describe using a "hash-matching" method and sentences shorter than eight words were

	excluded from the method. However, disputed that the eight-word-or-longer sentences were "unique." The cited paragraphs do not support the assertion that each eight-word-or-longer sentence is unique to any of the Asserted Works. And while the cited paragraphs state that sentences shorter than eight words were excluded from the hash-matching method "to guard against common short phrases," they do not contend that all remaining sentences are unique. Defendants also object to "Defendants' data" without identifying the specific defendant or data set about which the alleged fact applies. Such statements are unintelligible such that Defendants cannot properly respond, and so dispute said statements.
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.[2]	919. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that the cited paragraph from Dr. Shan's Opening Report describes a methodology he calls "Quartile-Matching." Disputed that this methodology "ensure[d] the match was to a full copy of the book." That characterization of Dr. Shan's methodology is not supported by Dr. Shan's analysis in the cited paragraph, as the quartile-based thresholds discussed in the cited paragraph, even if met, do not themselves establish that the undefined "Defendants'

[2] In a limited number of cases where a translated copy of the work was present, Dr. Shan manually verified the match using Google Translate. ECF 1814-1 (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.

	data" contained a full copy of the work. Indeed, under this methodology, up to 80% of each quarter could be missing. Also disputed that the sentences used in this methodology are "unique" as the cited paragraph does not support the assertion that every sentence is unique to an Asserted Work. Further disputed that the methodology established that "a large share" of the unique sentences in each quartile was present; this mischaracterizes Shan's methodology, which is tethered to particular thresholds he selected. Further, for the Bing index specifically, Shan's "full copy" quartile matching identified only 2 works; the vast majority (22) were identified under the lower 11% threshold. This demonstrates that the quartile-matching method — designed to ensure matches to "full copies" — failed for the overwhelming majority of Bing index matches, contradicting the claim that this method reliably identifies full books. C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) at 77–78. Thus, the disputed proposition is not supported by the cited evidence.
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.	920. **Disputed in Part – Not Supported by Cited Evidence.** Undisputed that the cited paragraphs from Dr. Shan's Opening Report describe a methodology involving metadata. However, disputed that OpenAI "deleted" the underlying Books3 data, as that characterization is not supported by the record evidence. OpenAI's corporate designee testified that, outside of the Books1 and Books2 datasets previously discussed, he was "not aware of any" other datasets being deleted and

	confirmed that OpenAI did not have a practice of deleting datasets. Books Ex. 62(a) (M. Trinh 30(b)(6) Tr. (Jan. 29, 2025)) at 59:7–60:8. Microsoft objects that Dr. Shan conflates Defendants' alleged acts by asserting "OpenAI ***or*** Microsoft" deleted LibGen, Books3, and the Bing Index. Nowhere does the cited evidence assert that Microsoft was in possession of, let alone deleted, the LibGen or Books3 datasets that Dr. Shan alleges OpenAI used to train certain of the models at issue. Further, the Bing Index is not a dataset that was ever used by OpenAI to train any model at issue, nor was it deleted by Microsoft. It is a living index of the internet that is updated regularly. Further, when asked directly whether he "actually [did] any comparison of the underlying text" for matches identified solely from his metadata methodology, Shan acknowledged: "For the second, you know, the amounts of metadata here, I did not use binary hash matching." C. Blythe Opp. Decl. Ex. 73 (S. Shan Tr.) at 99:8-24. And for at least some of the matches identified solely through his metadata methodology, Shan admitted that he "[did] not have the underlying files for those files." *Id.* at 101:20-21. He described his verification as checking whether quartile-matched books also appeared in Libgen metadata — not independently comparing underlying text against the datasets where data was allegedly deleted. *Id.* at 99:22-100:9. His verification was circular — he checked metadata against results from his other method, rather than independently confirming that

	the metadata accurately reflected the deleted underlying content. *Id.* at 100:18-102:1. Further, the disputed proposition is not supported by the cited evidence. Defendants further respond that it is undisputed that Plaintiffs are not moving for summary judgment based on alleged "copies" identified by Dr. Shan using an 11% threshold analysis or any "manually verified" translated works. However, disputed that Dr. Shan's efforts to purportedly "manually verif[y]" translated works using Google Translate establishes that the translated copies were in fact "copies" of Plaintiffs' Works, as no details of such manual verification are provided. Also disputed that Dr. Shan verified any "copy" of any Asserted Work within the meaning of copyright law, as that is a legal conclusion.
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.	921. **Disputed in Part.** Defendants respond as follows: Disputed that Dr. Shan "verified" the accuracy of the metadata and "confirm[ed]" that it accurately matched the underlying contents. That characterization is not supported by the cited paragraphs in Dr. Shan's Opening Report, which speak for themselves. For example, the cited paragraphs do not describe any "verification" process for Books3 or URL data, or any specific actions done to "verify" metadata accuracy. Moreover, Plaintiffs' characterization improperly elevates Dr. Shan's analysis into independent verification that the metadata itself was accurate, which it is not.

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.	922. **Undisputed.**
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.	923. **Disputed in Part – Improper Legal Argument.** Defendants respond as follows: Undisputed that S. Shan Decl. Ex. M shows two columns of text side by side, with the left column entitled "Plaintiff Book" and the right column entitled "Defendant Dataset," and the information at the top of the exhibit identifies the book as John Grisham's "The Firm." However, disputed to the extent Plaintiffs contend that the "corresponding matching record" is an unauthorized or unlawful "copy" under copyright law, as that is a disputed legal conclusion. The Court should disregard the disputed portion of Paragraph 923 because the assertion constitutes improper attorney argument and is conclusory. It thus requires no response.
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.	924. **Undisputed.**
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.	925. **Undisputed.**

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.	926. **Disputed in Part – Not Supported by Cited Evidence.** Defendants respond as follows: Undisputed that exact hash matching may fail to identify text containing differences from the comparison text. Further, Defendants dispute the statement as speculative and argumentative: the cited material does not quantify any OCR- or Bates-related effect on Dr. Shan's results.
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.	927. **Undisputed.**
928. Nonetheless, the match works with a high degree of fidelity, as shown by this example taken later in the book:  Shan Decl. Ex. M.	928. **Disputed – Not Supported by Cited Evidence.** Defendants respond as follows: The facts set forth in Paragraph 928 are disputed because they are not supported by the record. Specifically, Plaintiffs' assertion that Dr. Shan's matching methodology "works with a high degree of fidelity" is a disputed generalized characterization of Dr. Shan's methods, and is attorney argument as opposed to a statement of fact. Only one example is cited to support this assertion, which does not establish that Dr. Shan's methods worked "with a high degree of fidelity" across all data analyzed.
929. Plaintiffs have included on this hard drive a spreadsheet entitled "Class Plaintiffs MSJ Match Sheets"	929. **Undisputed**.

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).	930. **Disputed in Part.** Defendants respond as follows: Undisputed that Dr. Shan reports that he identified all 194 works through Quartile or Metadata Matching. However, Defendants dispute this fact's characterization of "Defendants' data" as it treats various distinct, unrelated data sets from two distinct parties as one set of "data" without identifying the specific defendant or data set about which the alleged fact applies. Such statements are unintelligible such that Defendants cannot properly respond, and so dispute said statements.
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.	931. **Disputed in Part.** Defendants do not dispute that ECF 1814-12 (S. Shan Decl. Ex. L), reports Quartile matches for 163 works and Metadata matches for 189 works. Defendants dispute this fact to the extent it treats various distinct, unrelated data sets from two distinct parties as one set of "data," without identifying the specific defendant or data set about which the alleged fact applies. Such statements are unintelligible such that Defendants cannot properly respond, and so dispute said statements.

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	932. OpenAI responds as follows: Undisputed that Dr. Shan makes the assertion that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 2 ("1 – Metadata") reports his purported

in OpenAI's Books3 download based on metadata. *See* Shan Decl. Ex. L Tab 2 (1-Metadata) Columns C and D.	metadata identifications for 187 works in LibGen metadata and 143 works in Books3 metadata. Microsoft defers to OpenAI.
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.*	933. OpenAI responds as follows: Undisputed that Dr. Shan makes the assertion that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 3 ("2 – Radford 2018 LibGen Download") reports Quartile matches for 96 works and lists corresponding visualization filenames in Column C. Microsoft defers to OpenAI.
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.*	934. **O**penAI responds as follows: Undisputed that Dr. Shan makes the assertion that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 4 ("3 – Mann Hallacy 2019-20 Download") reports 187 works identified by Quartile or Metadata Matching. The spreadsheet reports 95 Quartile-matched works and 187 metadata-matched works; Column D lists visualization filenames for the Quartile matches. Microsoft defers to OpenAI.
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.*	935. OpenAI responds as follows: Undisputed that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 5 ("4 – Books3") reports Quartile matches for 14 works in the produced Books3 material, metadata identifications for 143 works, and corresponding Quartile visualization filenames in Column D.

	Microsoft defers to OpenAI.
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.*	936. OpenAI responds as follows: Undisputed that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 6 ("5 – LibGen1 and LibGen2") reports Quartile matches for 124 distinct works across LibGen1 and LibGen2—80 in LibGen1 and 91 in LibGen2—and lists corresponding visualization filenames in Columns D and E. Microsoft defers to OpenAI.

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.*	937. OpenAI responds as follows: Undisputed that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 9 ("8 – In-Suit Models Training Data") reports 157 distinct works across the listed model-associated datasets, with model-specific totals of 131, 125, 49, 58, 50, and 97, and lists corresponding Quartile visualization filenames. Microsoft defers to OpenAI.

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.*	938. OpenAI responds as follows: Undisputed that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 7 ("6 – In Common Crawl datasets") reports Quartile matches for 71 works and lists corresponding visualization filenames in Column C. Microsoft defers to OpenAI.
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.*	939. Defendants do not dispute that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 8 ("7 – In GPT-Bot datasets") reports Quartile matches for 49 works, attributes 27 of those works to Microsoft's Project Mango downloads, and lists visualization filenames in Column C.
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.*	940. Defendants do not dispute that Dr. Shan asserts that ECF 1814-12 (S. Shan Decl. Ex. L), Tab 10 ("9 – other datasets") reports matches for 29 distinct works across the listed datasets and file locations, including 2 in the Bing Index and 26 in the "booksummaries" folder, and lists corresponding visualization filenames.

IX. Defendant OpenAI's Additional Facts Pursuant to Local Rule 56.1(b)

941. OpenAI's corporate designee testified that "the intention of all of the data that [OpenAI] collect[s] is for the purposes of training." C. Blythe Opp. Decl. Ex. 21 (I. Sohl 30(b)(1) Tr. (Dec. 10, 2025)) at 114:25–115:2.	
942. OpenAI sought to obtain "a diverse set of text over many topics." C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 44:3–44:7.	
943. OpenAI considered books another "natural topic of diverse material[.]" C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 44:17–45:9.	
944. Books comprised only one category of data, alongside data such as source code, math problems, legislative materials, and public web data. C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 44:17–45:9.	
945. OpenAI witnesses testified that OpenAI downloaded a "small fraction" of The Pile "for the purpose of evaluating a very small model on it." C. Blythe Opp. Decl. Ex. 42 (S. Jain 30(b)(1) Tr.) at 316:23–317:3; C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 214:14–21, 215:12-216:8 (OpenAI used a "small fraction" of The Pile to run "small-scale experiments" in 2021).	
946. OpenAI's "very small scale" experiments done on a portion of The Pile involved "[v]ery, very tiny models" that "required a very little amount of data to train." C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 215:12–19.	

947. These experiments were run in order to obtain "some basic understanding of the quality of The Pile datasets" relative to OpenAI's existing training data, which was "the full extent of that work." C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 215:12–25.	
948. OpenAI did not use any books from Books3 to train any In-Scope Model. C. Blythe Opp. Decl. Ex. 41 (A. Paino 30(b)(6) Tr.) at 214:14–216:10.	
949. OpenAI's corporate designee testified the data actually transferred to Microsoft was only a "subset" of GPT-3's training data. C. Blythe Opp. Decl. Ex. 78 (B. Gaffney 30(b)(6) Tr.) at 147:13–148:25.	
950. The later transfer of GPT-3 training data was for [redacted]. C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr.) at 186:5-12.	
951. The goal was to use the data to [redacted], not to evaluate the data. C. Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr.) at 189:20-23. [redacted] *Id.* at 187:3-17.	
952. Only three of GPT-3's five component datasets were transferred and the datasets did not include the Library Genesis–derived Books1 and Books2 datasets. Books Ex. 275 at OPCO_MDL_001653588.	

953. Project Taxi included the transfer of portions of the Bing Index data so that OpenAI could evaluate the value of the data for training out-of-scope "internal-only research models." C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 344:19–24, 345:10–19, 345:16–20; C. Blythe Opp. Decl. Ex. 22 (Q. Yuan 30(b)(1) Tr.) at 77:18–21; C. Blythe Opp. Decl. Ex. 62 (N. Ryder 30(b)(6) Tr. (Jan. 29, 2026)) at 66:1–4.	
954. OpenAI used the Bing Index data to train "internal-only research models for evaluation purposes." C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 345:10–13; *see also id.* at 445:17–18 ("Project Taxi was used for training, but not training a production model.").	
955. OpenAI did not include the portions of Bing Index data in production-model training. C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 345:16–20.	
956. OpenAI eventually deleted the data. C. Blythe Opp. Decl. Ex. 49 (I. Sohl 30(b)(6) Tr. (Dec. 9, 2025)) at 346:3–9.	
957. For the "book summarization project," OpenAI used a custom-built web portal to display very short snippets to contractors evaluating model generated summaries. C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) 126:19–127:23, 128:14–129:1.	

958. The project tested experimental training techniques by examining whether breaking difficult tasks into smaller, more readily evaluated subtasks could improve model training. *See* C. Blythe Opp. Decl. Ex. 35 at OPCO_MDL_000000048, -050, -059; C. Blythe Opp. Decl. Ex. 23 (J. Wu 30(b)(1) Tr.) 121:23–122:8.	
959. [redacted]. C. Blythe Opp. Decl. Ex. 70 ([redacted].) 182:21–184:1.	
960. [redacted]. *See* C. Blythe Opp. Decl. Ex. 70 ([redacted]) 182:21–184:1, 222:4–11.	
961. [redacted] G. Leonard Opp. Decl. ¶ 25.	

962. The majority of the Asserted Works are adult fiction trade books by well-known authors and published by the Big 5 publishers. G. Leonard Opp. Decl. ¶ 14.	
963. OpenAI's expert, economist Dr. Gregory Leonard, noted that traditionally published books such as the Asserted Works are substantially differentiated from, and thus experience little competitive effect from, self-published works. G. Leonard Opp. Decl. ¶¶ 15, 19.	
964. Self-published books are often limited to digital formats, and self-published authors do not receive an advance against royalties. G. Leonard Opp. Decl. ¶ 15.	
965. Self-published books do not have the extensive marketing resources and relationships that traditional publishers have built to reach target audiences. G. Leonard Opp. Decl. ¶ 15.	
966. Factors driving demand for self-published works are typically different than factors driving demand for traditionally published books. For example, author loyalty and reputation are strong drivers of the demand for books. G. Leonard Opp. Decl. ¶ 17.	
967. Similarly, publisher support is a driver of demand because large publishers' sales, marketing, and publicity teams promote the discovery of traditionally published titles. G. Leonard Opp. Decl. ¶ 18.	

968. Only about one quarter of readers find books by browsing or keyword searching. G. Leonard Opp. Decl. ¶ 18.	
969. Given the structural differences, self-published books are less likely to access similar drivers of demand and channels of discovery. G. Leonard Opp. Decl. ¶¶ 15, 19.	
970. In July 2023, Authors Guild CEO Mary Rasenberger wrote that, [redacted] C. Blythe Opp. Decl. Ex. 79 (FICTION-0166829) at FICTION-0166829.	
971. At a December 2023 Authors Guild Council meeting, CEO Mary Rasenberger stated [redacted] C. Blythe Opp. Decl. Ex. 80 (FICTION-0060987) at FICTION-0060993–94.	
972. Authors Guild Chief Legal Officer Kevin Amer testified [redacted]	

[redacted]. C. Blythe Opp. Decl. Ex. 82 (K. Amer 30(b)(1) Tr.) at 304:22–306:3.	
973. The Authors Guild [redacted] C. Blythe Opp. Decl. Ex. 1 (M. Rasenberger 30(b)(6) Tr. (Feb. 13, 2026)) at 54:23–55:10.	
974. Authors Guild CEO Mary Rasenberger testified that [redacted]. C. Blythe Opp. Decl. Ex. 1 (M. Rasenberger 30(b)(6) Tr. (Feb. 13, 2026)) 58:11–59:10.	
975. Authors Guild CEO Mary Rasenberger testified that [redacted] C. Blythe Opp. Decl. Ex. 1 (M. Rasenberger 30(b)(6) Tr. (Feb. 13, 2026)) 59:2–10.	
976. Plaintiff Michael Connelly testified that [redacted] C. Blythe Opp. Decl. Ex. 88 (M. Connelly 30(b)(6)/30(b)(1) Tr.) 156:1–9.	
977. Plaintiff Jonathan Franzen has said that [redacted] C. Blythe Opp. Decl. Ex. 89 (J. Franzen 30(b)(1) Tr.) 42:5–20.	
978. Jordan Usdan, the GM, Data Strategy and Ops at Microsoft AI, testified that Microsoft's agreement with	

[redacted] C. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 271:3–13.	
979. The [redacted]. C. Blythe Opp. Decl. Ex. 90 at OPCO_MDL_004600927, OPCO_MDL_004600931.	
980. The [redacted]. C. Blythe Opp. Decl. Ex. 90 at OPCO_MDL_004600930.	
981. In [redacted] C. Blythe Opp. Decl. Ex. 91 at OPCO_MDL_004649467-470.	
982. [redacted] C. Blythe Opp. Decl. Ex. 91 at OPCO_MDL_004649467-470.	
983. [redacted]	

[redacted] C. Blythe Opp. Decl. Ex. 92 at OPCO_MDL_004590656–57.	
984. [redacted]. C. Blythe Opp. Decl. Ex. 92 at OPCO_MDL_004590656.	
985. [redacted] C. Blythe Opp. Decl. Ex. 92 at OPCO_MDL_004590656.	
986. Plaintiffs' expert examined approximately 338 million outputs from 108 million conversation logs and identified reproductions of 30 or more words from just seven of Plaintiffs' works in only 31 conversations—a per-conversation regurgitation rate of 0.00003%. ECF 1735, OAI SUF 91; T. Berg-Kirkpatrick Opp. Decl. ¶ 17.	

X. <u>Defendant Microsoft's Additional Facts Pursuant to Local Rule 56.1(b)</u>

Pursuant to Local Rule 56.1(b), Microsoft incorporates by reference Facts 1-33 as stated in its Statement of Undisputed Material Facts In Support of Defendant Microsoft's Motion for Summary Judgment, filed on September 4, 2026. ECF 1694 ¶¶ 1-33. Microsoft understands that Plaintiffs will be responding to those facts in connection with their opposition to that motion, and therefore Microsoft does not repeat those facts here.

Dated: October 23, 2026

Respectfully submitted,

*/s/ Allison S. Blanco**

LATHAM & WATKINS LLP
Andrew M. Gass (*pro hac vice*)
andrew.gass@lw.com
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
Telephone: 415.391.0600

Sarang V. Damle
sy.damle@lw.com
Luke A. Budiardjo
luke.budiardjo@lw.com
1271 Avenue of the Americas
New York, NY 10020
Telephone: 212.906.1200

Elana Nightingale Dawson (*pro hac vice*)
elana.nightingaledawson@lw.com
555 Eleventh Street, NW, Suite 1000
Washington, D.C. 20004
Telephone: 202.637.2200

/s/ Joseph C. Gratz

MORRISON & FOERSTER LLP
Joseph C. Gratz (*pro hac vice*)
jgratz@mofo.com
Tiffany Cheung (*pro hac vice*)
tcheung@mofo.com
Caitlin Sinclaire Blythe (*pro hac vice*)
cblythe@mofo.com
425 Market Street
San Francisco, CA 94105
Telephone: 415.268.7000

Lily Li (*pro hac vice*)
YLi@mofo.com
755 Page Mill Road
Palo Alto, CA 94304
Telephone: 650.813.5600

/s/ Robert A. Van Nest

KEKER, VAN NEST & PETERS LLP
Robert A. Van Nest (*pro hac vice*)
rvannest@keker.com
R. James Slaughter (*pro hac vice*)
rslaughter@keker.com
Paven Malhotra
pmalhotra@keker.com
Michelle S. Ybarra (*pro hac vice*)
mybarra@keker.com
Reid P. Mullen (*pro hac vice*)
rmullen@keker.com
Nicholas S. Goldberg (*pro hac vice*)
ngoldberg@keker.com
633 Battery St.
San Francisco, CA 94111
Telephone: 415.391.5400

Attorneys for Defendant OpenAI

* All parties whose electronic signatures are included in this filing have provided consent in accordance with Rule 8.5(b) of the Court's ECF Rules and Instructions.

/s/ Lisa Simpson

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

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

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

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

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

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

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

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

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

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

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

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

Attorneys for Defendant Microsoft Corporation