UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

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

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

Pursuant to Local Civil Rule 56.1, Defendant Microsoft respectfully submits this Reply to Books Plaintiffs' Response to Rule 56.1 Statement of Undisputed Material Facts in Support of Microsoft's Motion for Summary Judgment. ECF No. 2213 (the "Statement").

Pursuant to this Court's Order dated October 7, 2026 (ECF No. 2092), Defendant Microsoft provided its response in column format and has attempted to streamline its response for the benefit of the Court. For the Court's convenience and to avoid further lengthening this response, Microsoft does not repeat any supporting authority for the issues below in every response. Instead, Microsoft states these objections plainly and incorporates the below authority into each of those responses.

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 Microsoft's summary judgment motion. Microsoft's designation of certain facts as "immaterial" does not concede that any fact not designated as such is material for purposes of adjudicating Microsoft's 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. *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 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 Microsoft introduces "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.

This reply also uses the following abbreviations and citations for briefs, Rule 56.1 statements, and record materials filed in connection with the parties' cross-motions for summary judgment in the Books Plaintiffs' case.

Citation	Document
Blythe Opp. Decl.	Declaration of Caitlin Sinclaire Blythe in Support of Defendants' Joint Response to Books Plaintiffs' Corrected Rule 56.1 Statement of Undisputed Material Facts in Support of Books Plaintiffs' Motion for Partial Summary Judgment
BP MSJ	Books' Plaintiffs' Memorandum of Law in Support of Motion for Partial Summary Judgment (ECF 1848)
BP Opp.	Books Plaintiffs' Opposition to Microsoft's Motion for Summary Judgment (ECF 2260)
BP SUMF ¶	Books Plaintiffs' Statement of Material Facts in Support of Books Plaintiffs' Motion for Partial Summary Judgment (ECF 1988)
BP Counter-SUMF ¶	Book Plaintiffs' Counterstatement of Material Facts in Opposition to Microsoft's Motion for Summary Judgment (ECF 2245)
Defs.' Resp. to BP SUMF ¶	Defendants' Response to Books Plaintiffs' Statement of Undisputed Material Facts in Support of Books Plaintiffs' Motion for Partial Summary Judgment (ECF 2237)
ECF 2050	Defendants' Notice of Motion and Rule 702 Motion to Exclude Expert Testimony
ECF 2055	Memorandum of Law in Support of Defendants' Rule 702 Motion to Exclude Expert Testimony
Gratz Decl.	Declaration of Joseph C. Gratz in Support of OpenAI's Motions for Summary Judgment

Citation	Document
Hurst Decl.	Declaration of Annette L. Hurst in Support of Defendant Microsoft's Motions for Summary Judgment Declaration of Annette L. Hurst in Support of Defendant Microsoft's Reply in Support of Microsoft's Motions for Summary Judgment
MSFT Books MSJ	Defendant Microsoft Corporation's Memorandum of Law in Support of Motion for Summary Judgment in Books Plaintiffs' Consolidated Cases (ECF 1692)
MSFT Books SUMF ¶	Defendant Microsoft Corporation's Rule 56.1 Statement of Undisputed Material Facts in Support of Motion for Summary Judgment in Books Plaintiffs' Consolidated Cases (ECF 1694)
Nelson Decl.	Declaration of Justin Nelson in Support of Books Plaintiffs' Motion for Partial Summary Judgment Declaration of Justin Nelson in Support of Books Plaintiffs' Opposition to Defendants Microsoft's and OpenAI's Motions for Summary Judgment
Simpson Decl.	Declaration of Lisa T. Simpson in Support of Defendant Microsoft's Oppositions to the Books and News Plaintiffs' Motions for Partial Summary Judgment

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

A. Fair Use (17 U.S.C. § 107)

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

Paragraph 1

1. A large language model ("LLM") is a machine learning deep neural network trained on vast quantities of text with the objective of predicting the next token in a sequence in order to generate natural language. Lafferty Decl. ¶¶ 16, 33, 77-78; *see also* Hurst Decl. Ex. 75 (Melton Dep. Tr.) 29:6-30:19, 9:20-10:16, 11:16-17:7; Ex. 112 (Song Dep. Tr.) 32:13-33:25, 9:8-10:21; Ex. 54 (Huynh Dep. Tr.) 95:23-96:4, 7:11-8:18, 15:3-7, 15:25-16:25; Ex. 81 (Nadella Dep. Tr.) 274:24-276:4, 277:3-23, 17:19-18:7, 25:16-22, 29:10-30:3; Ex. 115 (Suleyman Dep. Tr.) 35:9-36:8; Ex. 99 (Ribas Dep. Tr.) 22:2-25, 10:23-11:6, 11:23-12:3, 14:2-23, 17:9-18:10, 18:3-19:19.	
Class Plaintiffs' Response: Class Plaintiffs object to the phrase "vast quantities of text" as vague and ambiguous. Subject to and without waiving this objection, **Undisputed** that a large language model ("LLM") is "a machine learning deep neural network" trained on text with the objective of predicting the next token in a sequence. **Disputed** to the extent Microsoft contends that an LLM must be "trained on vast quantities of text." LLMs can be trained on comparatively small amounts of high-quality data like books. Microsoft researchers trained the phi-1 LLM on 7 billion tokens of pre-training data—roughly 42 times fewer than the ~300 billion OpenAI used to train GPT-3—and found that "[b]y intentionally selecting and generating high-quality data," they could "achieve state-of-the-art results on code-generation tasks with a much smaller model and less compute than existing approaches." Nelson Decl. Ex. 937 at 1, 4. A model trained on less data	**Microsoft's Reply:** Paragraph 1 is not genuinely disputed. **Undisputed** that a large language model ("LLM") is a machine learning deep neural network trained on text with the objective of predicting the next token in a sequence. Books Plaintiffs[1] seek to create a dispute of fact by arguing that (1) LLMs need not be trained on "vast" quantities of text, (2) only quality of training data matters, and (3) that generating natural language is not an objective of training LLMs. None of these purported disputes creates a genuine issue of material fact, because (1) the evidence cited by Books Plaintiffs does not controvert the fact as stated, (2) the response is improper legal argument, and (3) the response answers a different proposition than the one stated in the fact. *First*, Paragraph 1 describes how LLMs are trained; it does not set a minimum corpus size. Instead of controverting the

[1] Microsoft uses the term "Books Plaintiffs" throughout to refer to the Named Plaintiffs in the purported class action.

can likewise perform domain-specific tasks when trained on domain-specific data. Nelson Decl. Ex. 938 at 76:18–23. OpenAI's own records further confirm that data quality, not volume, drives model performance. OpenAI's Shantanu Jain stated that "if forced to pick just one, i'd much prefer a dataset of 1M documents that we trust over 5M we don't," Nelson Decl. Ex. 890 at -749, and an internal OpenAI document explained that "if you add 1% of a high-quality data you'll get much better downstream performance since the model will cover that and you'll get it at test time," Nelson Decl. Ex. 300 at -749. Further **Disputed** to the extent Microsoft contends that the object of next-token prediction is to generate novel natural language rather than to approximate the text on which the model was trained. OpenAI's former Chief Technology Officer Mira Murati rejected that framing, testifying that she did "not believe that these models were trained with this specific goal" of generating text that appears human-generated. Nelson Decl. Ex. 936 at 105:2-13; *see also* Zhao Decl. Ex. B ¶¶ 14, 26, 43; 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).	stated fact, Plaintiffs instead seek to add an additional fact to create a dispute: that smaller models designed for a specific task need less data than the GPT models at issue. A response that addresses a different proposition does not controvert the one stated. Further, the vagueness objection is improper legal argument. "Vast quantities of text" has a plain meaning, and Plaintiffs' own expert uses similar terminology. *See* ECF 1787 (Zhao Decl.) Ex. B, ¶ 54 (describing "voluminous training data"). Indeed, Books Plaintiffs' cited evidence is consistent with the stated fact. The phi-1 model discussed in Nelson Decl. Ex. 937 was pre-trained on 7 billion tokens, which, while less than that used to train the models at issue, does not refute that "vast quantities of text" are needed for training. Nelson Decl. Ex. 937 also states that because the model "has only 1.3B parameters trained with only 7B tokens, this restricts [the] model's capacity to manage more complex tasks[.]" Nelson Decl. Ex. 937 (Textbooks Are All You Need Microsoft Research Paper) at 14, 20. Similarly, the full quote from Ex. 890 says, "[redacted]" (emphasis added). Nelson Decl. Ex. 890 (S. Jain Dep. Ex. 99 – S. Jain and A. Konstantinidis messages (dated May 19, 2021)) at -749. *Second*, Books Plaintiffs' response includes improper legal argument and an improper recitation of facts regarding the relative value of data *quality*, which has no bearing on whether LLMs are trained on *vast quantities* of text. Such additional facts, which do not contradict the stated fact, should be disregarded. *Third*, the dispute "to the extent Microsoft contends that the object of next-token prediction is to generate novel natural

	language" also answers a proposition not stated in Paragraph 1. Whether the models were trained with the "specific goal" of producing text that appears human-generated does not controvert the fact that LLMs predict the next token in order to generate natural language. Ms. Murati further testified [redacted] Hurst Decl. Ex. 80 (M. Murati Tr.) at 108:4-10. [redacted]" *Id.* at 108:12-18. There is no dispute that training aims to predict the next token in a sequence, and natural-language generation is the result of that training. Microsoft's LLM expert agrees that "LLMs are deep learning systems designed to process and generate human language," LLMs are trained on "vast quantities of text, with the objective of predicting the next token in a sequence," and that "[t]hrough the pre-training process, the parameters of the neural network represent a vast amount of generalized information about the semantic structure, grammar, style, and other types of word co-occurrence patterns underlying the entirety of data it was trained on." ECF 1723 (Lafferty Decl.) ¶¶ 77-78. "[N]ext-token prediction provides a self-supervised learning signal that allows the model to learn grammar, syntax, factual associations, style, and other statistical regularities across the corpus." *Id.* ¶ 272. Similarly, Books Plaintiffs' expert Dr.

	Zhao agreed that in training "[u]ltimately, the model will learn the most common patterns in the grammar, syntax, and semantics of a given language." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 29:15-24.

Paragraph 2

2. Machine learning requires large amounts of data, and the more complex and varied the domain, the greater the volume and diversity of data needed for a model to learn patterns that generalize reliably to new inputs. Natural language is among the most complex of such domains. Lafferty Decl. ¶¶ 16, 41, 45, 52, 66, 70, 77, 81, 90, 97, 99-100.

Class Plaintiffs' Response: Class Plaintiffs object to the phrase "among the most complex of such domains" as vague and ambiguous, because Microsoft neither identifies the "domains" against which natural language is compared nor supplies any metric for "complex[ity]." Subject to and without waiving that objection, **Disputed** that "[n]atural language is among the most complex of such domains," and **Disputed** that the complexity of a domain determines the volume or diversity of data required to train a model. Zhao Decl. Ex. B ¶¶ 43, 45. **Further Disputed** that all machine learning necessarily requires large amounts of data to train. Admissible evidence contradicts this statement. As set forth in greater detail in response to Microsoft's Statement of Fact No. 1, certain smaller models trained on smaller amounts of curated data can outperform larger models trained on larger amounts of lower-quality data. **Further Disputed** because the training objective for LLMs is next-token prediction, which means adjusting the	**Microsoft's Reply:** Paragraph 2 is not genuinely disputed. **Undisputed** that the more varied a domain, the greater the volume and diversity of data a model needs; their response challenges only the role of "complexity." Books Plaintiffs also do not dispute that "generalization occurs." To the extent Books Plaintiffs purport to dispute Paragraph 2, such dispute is not genuine because the cited evidence does not specifically controvert the fact as stated. *First*, Books Plaintiffs purport to dispute that natural language is not "among the most complex of such domains" and that "the complexity of a domain" does not affect "the volume or diversity of data required to train a model." The cited evidence does not support either contention. Paragraphs 43 and 45 of Dr. Zhao's report say nothing about the complexity of natural language nor the amount of data needed for an LLM to produce it. ECF 1787 (Zhao Decl.) Ex. B, ¶¶ 43, 45. Indeed, Dr. Zhao's declaration supports the proposition. *Id.* Ex. A, ¶ 71 (Training LLMs on "a diverse array of textual genres and subject matter, foster[s] a more comprehensive understanding of language across various

parameters of the LLM so that its outputs match the training data as closely as possible. Zhao Decl. Ex. B ¶¶ 43, 45. Memorization and generalization are not opposites, and whatever generalization occurs depends on the model encoding the expression contained in its training data rather than acquiring an understanding of language independent of it. Zhao Decl. Ex. B ¶¶ 14, 27.	domains. … It is widely accepted that to achieve high quality answers to a wide-ranging diverse set of user queries, LLMs must be trained with a diverse set of training data.") (internal citation omitted). Further, Plaintiffs' vagueness objection is improper legal argument. Plaintiffs identify no ambiguity that prevents a response. *Second*, Plaintiffs again argue that not "all machine learning necessarily requires large amounts of data." This argument fails to create a genuine issue of material fact for the same reasons as stated in reply to the same argument in Response 1. *Third*, Plaintiffs' last argument that training means adjusting the parameters of the LLM so that its outputs match the training data as closely as possible and that generalization depends on the models encoding expression in the training data does not contradict the fact as stated: that models learn patterns from the training data that allows them to generalize to new inputs. Indeed, a paper cited by Dr. Zhao—"What Language Model to Train if You Have One Million GPU Hours?," Scao et. al., p. 1, 4, Nov. 2022 (available at: https://arxiv.org/pdf/2210.15424)– supports that diverse pretraining data "improves zero-shot generalization", which is an emergent capability in LLMs. ECF 1787 (Zhao Decl.) Ex. A, ¶ 71.

Paragraph 3

3. LLMs are a general-purpose technology capable of a wide variety of natural language processing tasks, such as summarization, translation, reasoning, content generation, and question answering. Lafferty Decl. ¶¶ 16, 33, 36, 45, 77-78; *see also* Hurst Decl. Ex. 112 (Song Dep. Tr.) 32:13-33:25; Ex. 54 (Huynh Dep. Tr.) 95:23-96:4, 7:11-8:18, 15:3-7, 15:25-16:25, 17:14-22; Ex. 81 (Nadella Dep. Tr.) 274:24-276:4, 277:3-23;

Ex. 115 (Suleyman Dep. Tr.) 35:9-36:8; Ex. 99 (Ribas Dep. Tr.) 18:3-19:19, 22:2-25; Ex. 119 (Templeton 30(b)(6) Dep. Tr. (Vol. 2, dated Jan. 21, 2026)) 407:24-408:9; Ex. 118 (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 13:14-16, 14:14-20, 15:2-12, 16:18-19:10, 22:1-14, 22:25-23:15, 24:2-25:25, 26:14-27:16, 221:6-222:20.	
Class Plaintiffs' Response: Undisputed that LLMs are capable of natural language processing tasks, such as summarization, translation, content generation, and question answering. **Disputed** that LLMs are capable of "reasoning." LLMs are statistical prediction machines that mimic human reasoning. Zhao Decl. Ex. B ¶¶ 14, 16. According to Microsoft Director of Applied Science Brent Hecht, LLMs "don't reason," and though an LLM's "pattern replication" can "approximate[] reasoning," there are "major ways in which it [does] not approximate it very well." Nelson Decl. Ex. 939 at slide 6; Nelson Decl. Ex. 940 at 300:6-25. Microsoft's technical expert John Lafferty has likewise acknowledged that "the ability to generalize far outside the training distribution remains a significant obstacle[.]" Nelson Decl. Ex. 941 at 3. **Further Disputed** to the extent Microsoft invokes the phrase "general-purpose technology" in a misleading way, as an undisputed economic definition rather than a description of technology. Dkt. 1875-1 McDonald Rebuttal Rep. at 47 (describing the General Purpose Technology Framework is drawn from the economics of growth and innovation literature and is used to explain patterns of aggregate productivity growth but is not meant to be used as an analytical framework for evaluating effects on markets or industries).	**Microsoft's Reply:** Paragraph 3 is not genuinely disputed. **Undisputed** "that LLMs are capable of natural language processing tasks, such as summarization, translation, content generation, and question answering," and Books Plaintiffs concede that LLMs "mimic human reasoning" and that an LLM's "pattern replication" can "approximate[] reasoning." However, Plaintiffs seek to create a dispute of fact that LLMs are not capable of *human* "reasoning." Paragraph 3 does not say "human" reasoning nor does it anthropomorphize the general-purpose nature of the LLM technology or imply human psychology. Reasoning in this sense is simply an objective empirical description of a natural language processing function that LLMs can perform by engaging in multiple steps in a chain of inferences to obtain a result. As such, Books Plaintiffs do not specifically controvert the fact as stated. Plaintiffs also dispute Microsoft's use of the phrase "general-purpose technology" to the extent the phrase is used to mean "an undisputed economic definition rather than a description of technology." Paragraph 3 uses the phrase descriptively, as the rest of the sentence makes clear: LLMs are "capable of a wide variety of natural language processing tasks." Books Plaintiffs improperly cite the definition of "general-purpose technology" used by an expert in a different context who is not in this case. Specifically, Books Plaintiffs cite the McDonald Rebuttal Report which uses "general-purpose technology" in relation

	to an economic framework used "to explain patterns of aggregate productivity growth." It is improper for Books Plaintiffs to rely on an expert report from a separate litigation in the MDL and, even if it were proper, the cited evidence does not dispute that LLMs can perform a wide variety of tasks, and therefore does not specifically controvert the fact as stated. Microsoft uses the phrase in the technical sense to describe a technology that can be used for a wide variety of tasks. *See* ECF 1723 (Lafferty Decl.) ¶ 16. As such, Books Plaintiffs do not specifically controvert the fact as stated.

Paragraph 4

4. The process of training an LLM involves:	
Class Plaintiffs' Response: This umbrella clause is **Disputed** to the extent that "training" refers to anything other than pretraining a large language model, and is further **Disputed** to the extent it suggests the ensuing sub-paragraphs contain a complete description of the entire model training process. LLM training includes additional phases that the ensuing sub-paragraphs do not describe, including mid-training and post-training. Nelson Decl. Ex. 352 at -422; *see also* Nelson Decl. Ex. 947 at -870 (members of OpenAI's mid-training team describing accuracy goals for mid-training).	**Microsoft's Reply: Undisputed** that the process of training an LLM involves the steps set out in Paragraphs 4(a)–(e). Plaintiffs' response goes beyond the scope of Paragraph 4, which states only that the training process "involves" the listed steps. Paragraph 4 does not purport to "contain a complete description of the entire model training process." That paragraph 4 does not detail every step of every stage of what could be considered "training" does not create a genuine issue of material fact. As such, Books Plaintiffs do not specifically controvert the fact as stated.
a. **Acquiring a large volume of text to assemble a training corpus—typically by crawling and scraping publicly accessible data on the internet in a manner that has existed before LLMs were invented;** Lafferty Decl. ¶¶ 69, 79, 89-91, 100; Tucker Decl. ¶¶ 117-122; *see also* Hurst Decl. Ex. 104	

(Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 32:5-33:14, 48:20-24, 66:23-67:10, 95:10-23; Ex. 130 (Yuan Dep. Tr.) 12:11-14:13, 28:6-12, 8:6-25, 10:9-23.	
Class Plaintiffs' Response: Disputed. Admissible evidence contradicts Microsoft's statement. First, LLMs can be trained without using "vast quantities of text" as set forth in greater detail in response to Microsoft's Statement of Fact No. 1. Second, there was no preexisting practice that permitted the scraping and crawling of content that would then be used to compete against the creators or owners of that content, or in violation of a website's terms of use. Zhao Decl. Ex. B ¶ 100. Third, neither crawling nor scraping of publicly available training data is a "typical" way to acquire training data. *See* Zhao Decl. Ex. C ¶ 36. Rather, the training corpus for an LLM can and does include sources other than web data obtained from crawling. For example, one can license content from rightsholders or hire humans to write the content needed for pretraining data. Lasinski Decl. Ex. A ¶¶ 98–110. Moreover, before pirating millions of books from Library Genesis, OpenAI repeatedly priced and budgeted for lawful acquisition. *See* Nelson Decl. Ex. 332 at -638, -637; Nelson Decl. Ex. 333 at -640, -639; Nelson Decl. Ex. 345; Nelson Decl. Ex. 293 at -989; Nelson Decl. Ex. 321; Nelson Decl. Ex. 698 at -870.	**Microsoft's Reply:** Paragraph 4(a) is not genuinely disputed. **Undisputed** that training an LLM involves "[a]cquiring a large volume of text to assemble a training corpus," and Books Plaintiffs concede that "the training corpus for an LLM can and does include . . . web data obtained from crawling." Book Plaintiffs seek to add additional facts to create a dispute regarding (1) the required size of the training corpus and (2) whether crawling and scraping publicly available data from websites to obtain data is "typical." That does not make this fact disputed. Books Plaintiffs do not specifically controvert the fact as stated and the response contains improper legal argument. Books Plaintiffs' argument that LLMs can be trained without using "vast quantities of text" fails to create a genuine issue of material fact for the same reasons as stated in Microsoft's Reply to the same argument in Paragraph 1. Books Plaintiffs do not dispute that training corpora include crawled web data but instead assert that crawling is not a "typical" means of acquiring training data, which does not create a genuine factual dispute because it is unsupported by the cited evidence and contradicts Books Plaintiffs' own expert, who testified that "many of these models are trained just from data scraped from the internet," (Hurst Decl. Ex. 370 (B. Zhao Tr.) at 81:18-19), and agreed that generally "the creator of an LLM gathers text, material and other content, typically in massive quantity and from a variety of sources," (*id.* at 28:4-29:9).

	Further, the portion of the response regarding whether there was a "preexisting practice that permitted scraping and crawling of content that would then be used to compete against the creators or owners of that content, or in violation of a website's terms of use" is improper legal argument about authorization, not a factual dispute about how training corpora are assembled. Further, this argument is not supported by the cited evidence. Zhao Declaration Ex. B, ¶ 100 says nothing about web crawling or website terms of use. *See* ECF 1787 (Zhao Decl.) Ex. B ¶ 100. Similarly, Zhao Declaration Ex. C ¶ 36 does not address whether crawling is a typical means of acquiring training data. *Id*. at Ex. C ¶ 36. Books Plaintiffs' assertion that one can "hire humans to write the content needed for pretraining data" finds no support in the paragraphs they cite: Lasinski Declaration Ex. A ¶¶ 98-110 discuss options to purchase, scan, or license books, but do not discuss hiring humans to write pretraining data. *See* ECF 1825 (Lasinski Decl.) Ex. A ¶¶ 98-110.
b. **Processing the textual data into a numerical format that can be used by a neural network—i.e., tokenization, whereby readable text is broken into unique "tokens" comprised of short words, segments of longer words, or groupings of characters, and then integers are assigned to each token;** Lafferty Decl. ¶¶ 105-114; *see also* Hurst Decl. Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 19:12-20:18, 176:22-177:23; Ex. 117 (Sutskever Dep. Tr.) 141:8-142:2, 19:15-19, 21:12-22:7, 26:17-29:8; Ex. 117.A (Sutskever Dep. Ex. 12 - Language Models Are Few-Shot Learners GPT-3 Research Paper) 8-11.	
Class Plaintiffs' Response: Undisputed that OpenAI converts language data into tokens. **Disputed** to the extent Microsoft contends that tokenization transforms the	**Microsoft's Reply:** Paragraph 4(b) is not genuinely disputed. **Undisputed** that "OpenAI converts language data into tokens," and their expert agrees that "[t]okenizers

underlying expression into something other than the text itself. As Class Plaintiffs' expert Ben Zhao explained: "Tokenizers are the statistical models which convert 'bytes' (the format in which text data is stored on a computer) to a set 'vocabulary' of numbers which the LLM can accept as input." Zhao Decl. Ex. B ¶ 35. In his deposition, OpenAI's CEO Sam Altman stated that "the standard byte pair encoding tokenization methods we have used … are reversible." Nelson Decl. Ex. 944 at 116:11-13.	are the statistical models which convert 'bytes' … to a set 'vocabulary' of numbers which the LLM can accept as input." ECF 1787 (Zhao Decl.) Ex. B, ¶ 35. To the extent Books Plaintiffs purport to dispute Paragraph 4(b), their response relies on improper legal argument and fails to specifically controvert the fact as stated. Whether tokenization "transforms the underlying expression into something other than the text" is improper legal argument directed to the conclusion of the first factor fair use inquiry, and does not dispute the technical process of tokenization that Paragraph 4(b) describes as a necessary step of translating human language into machine readable data. Further, the cited evidence does not specifically controvert the fact as stated. Dr. Zhao confirms exactly what Paragraph 4(b) describes, characterizing tokenization as turning text into integers. ECF 1787 (Zhao Decl.) Ex. B, ¶ 35. Plaintiffs' quotation of Mr. Altman also does not controvert the stated fact because reversibility of a token-to-integer mapping is consistent with Paragraph 4(b), which states that integers are assigned to each token. Nelson Decl. Ex. 944 (S. Altman Tr.) at 116:11-13.
c. **Embedding the tokens in high-dimensional, mathematical space, where tokens are represented by numerical vectors that point toward and away from each other based on relationships among the tokens gleaned from numerous contexts across the entirety of the training corpus;** Lafferty Decl. ¶¶ 33, 115-123; *see also* Hurst Decl. Ex. 117 (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 53:3-17, 67:9-15, 141:8-142:2; Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 50:10-19.	

Class Plaintiffs' Response: Undisputed that tokens are represented by numerical vectors reflecting relationships among tokens. **Disputed** to the extent Microsoft contends that those relationships are properties of the training corpus in the aggregate rather than of the individual works within it, as set forth in greater detail in response to Microsoft's Statements of Fact Nos. 10, 11, and 20.	**Microsoft's Reply:** Paragraph 4(c) is not genuinely disputed. **Undisputed** that tokens are embedded in high-dimensional, mathematical space, where they are represented by numerical vectors that point toward and away from each other based on relationships among the tokens. Book Plaintiffs assert that those relationships are due to properties of individual works instead of the training corpus in the aggregate. This assertion does not create a genuine factual dispute because it is unsupported by the cited cross-references. Plaintiffs admit in response to Paragraph 12 that LLMs "encode generalized statistical relationships of language ...." Further, Plaintiffs' own expert agreed that "the model will learn the most common patterns in the grammar, syntax and semantics of a given language" (Hurst Decl. Ex. 370 (B. Zhao Tr.) at 29:15-24) and testified that "most analyses show that there is no easy mapping between specific model weights or parameters and certain concepts or features." *Id.* at 354:11-16. Plaintiffs rely solely on cross-references to their responses to Paragraphs 10, 11, and 20. The cross-referenced material does not specifically controvert the fact as stated, because it concerns alleged memorization or reproduction of particular passages, not how token embeddings are derived from contexts across the training corpus.

d. **Processing those vectors through transformer layers repeatedly makes small adjustments to the model "parameters," a set of numerical model weights that the model uses to predict the next token in a sequence of text;** Lafferty Decl. ¶¶ 137-147; *see also* Hurst Decl. Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 18:16-19:4, 176:24-177:23; Ex. 101 (Rossett Dep. Tr.) 67:25-68:9, 8:1-9, 8:24-9:9, 17:6-18:4, 18:22-23:2; Ex. 117 (Sutskever Dep. Tr.) 141:8-142:2; Ex. 117.A (Sutskever Dep. Ex. 12 - Language Models Are Few-Shot Learners GPT-3 Research Paper) 2-8.

Class Plaintiffs' Response: Undisputed that the models at issue in this case encode numerical values in a network of parameters or weights. **Disputed** to the extent Microsoft suggests that "[p]rocessing those vectors through transformer layers repeatedly" transforms the underlying expression into something other than the text itself, as set forth in greater detail in response to Microsoft's Statement of Fact No. 4(b). Further **Disputed** to the extent Microsoft's phrasing suggests that the model adjusts itself. As Dr. Zhao explained, the "training process is simply a computer program that provides the model with input text from the training data, records whether the model's prediction is correct, measures the degree of error if not, and then applies basic calculus to quantify the contributions to the error and alter the model parameters such that the prediction is accurate." Zhao Decl. Ex. B ¶ 58.	**Microsoft's Reply:** Paragraph 4(d) is not genuinely disputed. **Undisputed** that the models at issue in this case encode numerical values in a network of parameters or weights the model uses to predict the next token in a sequence of text, or that the processing of vectors through transformer layers repeatedly causes small adjustments to the model's parameters. Book Plaintiffs' response instead seeks to add an additional fact regarding whether "the model adjusts itself," which is not included in the stated fact. That does not make this fact disputed as stated. Plaintiffs' response that the models do not encode expression into something other than the text itself does not create a genuine dispute because the response is improper legal argument that goes beyond the technical scope of Paragraph 4(d) and should be ignored by the Court.
e. **Periodically during training, the model will be presented with data it has not seen during training in order to measure the model's generalization ability. Training continues as long as the model's performance on the unseen data continues to improve. Once the model begins to perform well on training data but begins to lose accuracy on testing data, the training stops.** Lafferty Decl. ¶ 71.	
Class Plaintiffs' Response: Disputed insofar as the statement asserts that the objective of LLM training is generalization. The training objective is next-token prediction, which means adjusting the parameters of the LLM so that its outputs match the training data as closely as possible, and generalization is at most a hoped-for outcome rather than the objective, as set forth in greater detail in response to Microsoft's Statement of Fact No. 9. Zhao Decl. Ex. B ¶¶ 14, 16, 25-27; Zhao Decl. Ex. C ¶ 21;	**Microsoft's Reply:** Paragraph 4(e) is not genuinely disputed. **Undisputed** that, during training, the model will be presented with data it has not seen during training in order to measure the model's generalization ability. Book Plaintiffs' response instead seeks to add additional facts and make legal arguments to negate this general proposition regarding LLM training. That does not make this fact disputed as stated.

Nelson Decl. Ex. 50(a) at 175:6-176:10 (OpenAI's 30(b)(6) witness on LLM training topics: "For all of the models we are talking about, they were trained with next-token prediction loss.").	
Further Disputed to the extent Microsoft contends Defendants effectively withheld data "during training in order to measure the model's generalization ability." As Class Plaintiffs' expert Ben Zhao explained, "because OpenAI has not properly deduplicated or cleaned its training data—and given its admitted issues with contamination for in-scope models such as GPT-3—there is no basis to conclude that the use of such 'test' sets during training is reliably measuring generalization." Zhao Decl. Ex. B ¶ 70.	
Further Disputed that test loss rather than train loss is the operative metric. OpenAI's Alex Paino described "training loss" as "technically the mathematical function you're optimizing," Nelson Decl. Ex. 946 at 62:11-18, and as Dr. Zhao explains, "measurements of validation loss against supposedly held out data [are] not a constraint of the actual model training process," because the "loss function measured on the training data is the metric which is actually used to modify the model's parameters (and thus the resulting model)," Zhao Decl. Ex. C ¶¶ 21, 31; *see also* Nelson Decl. Ex. 945 (explaining most benchmarks for large language models are fully public and their questions are widely mirrored across the internet, so there is "systematic risk that models were trained on the very data used to evaluate them.").	
Further Disputed that "the training stops." There are additional phases of LLM training, including mid-training	*First*, Book Plaintiffs' cited evidence does not controvert the fact as stated. The phrase quoted from OpenAI's Alex Paino—"technically the mathematical function you're optimizing"—comes from [redacted] Nelson Decl. Ex. 946 (A. Paino 30(b)(6) Tr.) at 62:11-18. That testimony [redacted] which is exactly what Paragraph 4(e) states. Likewise, the 30(b)(6) testimony Plaintiffs cite explains that [redacted] Nelson Decl. Ex. 50(a) (N. Ryder 30(b)(6) Tr.) at 175:18-21. The testimony that the models "were trained with next-token prediction loss," *id.* at 175:22-176:1, is consistent with Paragraph 4(e). Further, Dr. Zhao's contamination opinions do not controvert that models are evaluated on held-out data, and Nelson Decl. Ex. 945 is a third-party study of contamination in public benchmarks, not the held-out validation data Paragraph 4(e) describes. *Second*, the "mid-training" and "post-training" exhibits do not create a dispute of material fact as they describe later adjustments to an already-pretrained base model. Nelson

and post-training. Nelson Decl. Ex. 352 at -422; *see also* Nelson Decl. Ex. 947 at -870 (members of OpenAI's mid-training team describing accuracy goals for mid-training).	Decl. Ex. 352 (OPCO_MDL_004740422) at -422; Nelson Decl. Ex. 947 (OPCO_MDL_000481870) at -870. *Third*, Books Plaintiffs' characterization of generalization as a "hoped-for outcome" and their cross-reference to their response to Paragraph 9 are improper argument rather than evidence.

Paragraph 5

5. Training an LLM requires an enormous volume of text. Publicly disclosed training-set sizes for frontier models rose from approximately 2 trillion tokens in 2023 to approximately 18 trillion tokens by late 2024. Lafferty Decl. ¶¶ 91-92; Hurst Decl. Ex. 215 (OPCO_MDL_004424223) at -226; Ex. 197 (MSFT_AI_MDL_000221039) at -042.	
Class Plaintiffs' Response: Disputed that "[t]raining an LLM requires an enormous volume of text." As set forth in greater detail in response to Microsoft's Statement of Fact No. 1, LLMs can be trained without using "an enormous volume of text." Smaller LLMs typically do not require trillions of tokens. Nelson Decl. Ex. 948 (describing an LLM that was trained on 300 billion tokens). Otherwise, **Undisputed.**	**Microsoft's Reply:** Paragraph 5 is not genuinely disputed. **Undisputed** that "[p]ublicly disclosed training-set sizes for frontier models rose from approximately 2 trillion tokens in 2023 to approximately 18 trillion tokens by late 2024." Book Plaintiffs attempt to create a dispute that "[t]raining an LLM" does not require "an enormous volume of text." For the reasons stated in Reply to Paragraphs 1, 2, 4(a), and 9, there is no genuine dispute that frontier LLMs, like those at issue in this case, require a "large," "enormous," or "vast" quantity of training data. Adding an additional fact about irrelevant "[s]maller LLMs" that are not the frontier models Paragraph 5 addresses does not make this fact disputed. Further, this fact is not genuinely disputed because the evidence cited by Books Plaintiffs is not admissible. Nelson Decl. Ex. 948 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.

Paragraph 6

6. The GPT models at issue were pre-trained on corpora of this scale: GPT-3 on approximately 300 billion tokens; GPT-4 and GPT-3.5 Turbo on approximately [redacted] tokens (epoch weighted); and GPT-4o on approximately [redacted] tokens (epoch weighted). Lafferty Decl. ¶¶ 91-92, 94-97; Hurst Decl. Ex. 196 (MSFT_AI_MDL_000205724); Ex. 215 (OPCO_MDL_004424223) at -226; Ex. 197 (MSFT_AI_MDL_000221039) at -042; Ex. 96 (Radford Dep. Tr.) 20:25-23:3; Ex. 227 (Paino Decl.).	
Class Plaintiffs' Response: Undisputed that it is reported that the GPT models at issue were pre-trained on corpora of this scale. **Disputed** to the extent Microsoft contends that those reported figures represent distinct tokens of data. In the paper titled "Language Models are Few-Shot Learners," OpenAI explained that "during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently," Nelson Decl. Ex. 162 at -660, and that "[a]s a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once," *Id.* at -661. OpenAI likewise told Microsoft that it [redacted] of its high-quality datasets "so that they occurred [redacted] of the time during training, [redacted] with the result that [redacted] Nelson Decl. Ex. 697 at -056.	**Microsoft's Reply:** Paragraph 6 is not genuinely disputed. **Undisputed** "that the GPT models at issue were pre-trained on corpora of this scale" or that the number of tokens (epoch weighted) cited were in the pre-training datasets for each of the named models. However, Book Plaintiffs seek to create a dispute of fact that the figures in Paragraph 6 do not "represent distinct tokens of data." But Paragraph 6 does not assert that the figures represent distinct tokens; for GPT-4, GPT-3.5 Turbo, and GPT-4o, it expressly states that the figures are "epoch weighted." The referenced research paper confirms that GPT-3 was trained "for 300 billion tokens," with some datasets seen more than once, Nelson Decl. Ex. 162 (Language Models Are Few-Shot Learners GPT-3 Research Paper) 11, 13, which is consistent with Paragraph 6.

Paragraph 7

7. Training an LLM also requires a wide diversity of text. Model developers draw training data from a broad range of sources, including web data, software code, books, encyclopedic and scientific sources, forum posts, and transcripts. Lafferty Decl. ¶¶ 34, 89-91, 93-97; Hurst Decl. Ex. 117 (Sutskever Dep. Tr.) 26:16-21, 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; Ex. 196 (MSFT_AI_MDL_000205724); Ex. 215 (OPCO_MDL_004424223) at -226; Ex. 197 (MSFT_AI_MDL_000221039) at -042; Ex. 198 (MSFT_AI_MDL_000221643); Ex. 213 (OPCO_MDL_001938121).	
Class Plaintiffs' Response: Disputed that "[t]raining an LLM also requires a wide diversity of text" in all circumstances, as set forth in greater detail in response to Microsoft's Statements of Fact Nos. 1 and 2. The assertion also presupposes a metric for quantifying training-data diversity, and OpenAI's own LLM-training expert is "unaware of research that would quantify diversity." Nelson Decl. Ex. 949 at 307:2-11. OpenAI's internal assessments turned on quality rather than breadth: OpenAI rated the quality of "Commoncrawl (filtered)" as [redacted] while rating its "Libgen" books dataset [redacted] and ran the [redacted] against Common Crawl's [redacted] Nelson Decl. Ex. 273 at -OP524. **Undisputed** only that model developers have drawn training data from the categories of sources Microsoft lists.	**Microsoft's Reply:** Paragraph 7 is not genuinely disputed. **Undisputed** "that [t]raining an LLM also requires a wide diversity of text" or that "[m]odel developers draw training data from a broad range of sources, including web data, software code, books, encyclopedic and scientific sources, forum posts, and transcripts." However, Plaintiffs seek to create a dispute of fact that training an LLM does not require "a wide diversity of text" "***in all circumstances***" and that diversity cannot be quantified. That purported dispute does not create a genuine issue of material fact because it fails to specifically controvert the fact as stated. The cited evidence does not specifically controvert the fact as stated. Plaintiffs truncate Dr. Callison-Burch's testimony, which was that he was "unaware of research that would quantify diversity in as useful a fashion as I would need to explain that opinion," Nelson Decl. Ex. 949 (C. Callison-Burch Tr.) at 307:2-11. That testimony concerns the absence of a metric for an opinion he holds about diversity; it does not state that diversity is unnecessary. He then explained his opinion "by example," testifying that a model trained on the Common Pile proved "insufficiently diverse because of the sourcing of that material being limited to public domain,

	freely licensed, et cetera, data," *id.* at 307:12-308:19, and that "using this massive and diverse dataset was necessary," *id.* at 305:23-306:8. Further, Paragraph 7 is supported by Plaintiffs' own expert. Dr. Zhao opines that books are valuable to LLM training in part because they provide "expressively varied text that enable LLMs to produce text in a wide-array of genres," ECF 1787t (Zhao Decl.) Ex. A ¶ 12(a)(iv), and he agreed that it is true that LLM creators gather content "from a variety of sources," Hurst Decl. Ex. 370 (B. Zhao Tr.) at 28:4-29:9.

Paragraph 8

8. The internet is the only practicable source of text at a scale sufficiently voluminous and varied to train a general-purpose LLM. Lafferty Decl. ¶¶ 45, 90-91; Usdan Decl. ¶ 10; *see also* Hurst Decl. Ex. 99 (Ribas Dep. Tr.) 80:3-81:23; Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 48:20-24, 67:2-6; Ex. 117 (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; Ex. 130 (Yuan Dep. Tr.) 28:6-25, 12:11-14:13.	
Class Plaintiffs' Response: Disputed. Admissible evidence contradicts Microsoft's statement. Microsoft's statement wrongly presupposes a strict dichotomy between the internet and other non-internet sources. In reality, model developers can use both internet and non-internet sources and thus reduce reliance on the internet. For instance, they can hire humans to write training data or acquire large volumes of books through licensing deals. *See* Lasinski Decl. Ex. A ¶¶ 98–110. In addition, the internet is not "singularly" suited to this purpose. As Dr. Zhao explained, "few online sources provide coherent, curated long form text similar to that of published books," which is "critical to performance on LLM tasks that require longer contexts." Zhao Decl. Ex. A ¶ 66.	**Microsoft's Reply:** Paragraph 8 is not genuinely disputed. **Undisputed** that model developers must use internet sources to train LLMs; their response asserts only that developers "can use both internet and non-internet sources and thus reduce reliance on the internet." Plaintiffs seek to create a dispute of fact that non-internet sources are a practicable alternative to the internet for text at the scale and variety needed to train a general-purpose LLM. But the evidence cited by Plaintiffs does not contradict the fact as stated. Lasinski Declaration Ex. A ¶¶ 98-110 discuss options to purchase, scan, or license books; those paragraphs do not discuss hiring humans to write training data and do not show that any non-internet source

	could supply text at the necessary scale. The same report quotes Microsoft's designee testifying that "without the access to the public web, … it'd be impossible to build … any large language model." *Id.* ¶ 98 n.482. Further, Books Plaintiffs' expert Dr. Zhao's statement regarding the use of published books as training data (ECF 1787 (Zhao Decl.) Ex. A ¶ 66) does not refute that a training corpus of the necessary scale can practicably be assembled without the internet. Plaintiffs cite no evidence of any general-purpose LLM trained without internet data. Indeed, Dr. Zhao testified that "many of these models are trained just from data scraped from the internet." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 81:18-19.

Paragraph 9

9. LLM training requires an enormous volume and a wide diversity of textual data because the fundamental objective of training is generalization—producing a model capable of responding to new, previously unseen inputs across an effectively unlimited range of contexts and tasks. Lafferty Decl. ¶¶ 16, 17, 34, 90-91, 98, 154-156.	
Class Plaintiffs' Response: Disputed. First, as set forth in greater detail in response to Microsoft's Statement of Fact No. 1, LLMs can be trained without using "an enormous volume" or "wide diversity of textual data." Second, the training objective for LLMs is next-token prediction, which means adjusting the parameters of the LLM so that its outputs match the training data as closely as possible. Zhao Decl. Ex. B ¶¶ 14, 25-27; Nelson Decl. Ex. 50(a) at 175:6-176:10 (OpenAI's 30(b)(6) witness on LLM training topics: "For all of the models we are talking about,	**Microsoft's Reply:** Paragraph 9 is not genuinely disputed. **Undisputed** that LLMs "will exhibit some amount of generalization" (ECF 1787 (Zhao Decl.) Ex. B ¶ 26; Hurst Decl. Ex. 370 (B. Zhao Tr.) at 338:3-13) or that LLMs are capable of responding to new, previously unseen inputs across an effectively unlimited range of contexts or tasks. Book Plaintiffs again attempt to create a dispute that LLM training does not require "an enormous volume and wide diversity of text." For the reasons stated in Reply to Paragraphs 1, 2, & 4(a), there is no genuine dispute that

they were trained with next-token prediction loss."); *see also* Nelson Decl. Ex. 298 at -468. Third, generalization can be a hoped-for outcome of model training, but it is not the training objective. As Dr. Zhao explains, the models "predict[] text in the distribution of the training corpus[,] [r]ather than engaging in a generalized process of learning language," Zhao Decl. Ex. B ¶ 16, and the "loss function is the quantitative metric which controls parameter updates during model training[,] . . . measure[ing] how accurately the model measures its training data, not human language as a whole," Zhao Decl. Ex. C ¶ 21. Statements from OpenAI employees confirm as much. James Betker, an OpenAI employee who has "trained a lot of generative models," wrote that LLMs "are truly approximating their datasets to an incredible degree," and that "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." Zhao Decl. Ex. B ¶ 16; Nelson Decl. Ex. 709. Jerry Tworek, a former OpenAI researcher and Vice President of Research, likewise stated that at OpenAI, "[w]e train our networks to memorize the training data" and "that's their objective." Nelson Decl. Ex. 274 at -365. Memorization and generalization are not opposites, and to the extent LLMs generalize at all they do so by encoding the expression contained in their training data. Shan Decl. Ex. D ¶¶ 14, 38; Zhao Decl. Ex. B ¶ 41 (the view that "models do not memorize but rather 'learn' and 'generalize' as humans do is fundamentally inconsistent with the mathematical process at the heart of LLM training: lossy compression").	LLMs of the scale of those at issue in this case, require a "large," "enormous," or "vast" quantity of varied training data. Books Plaintiffs' remaining response adds additional facts that do not controvert the stated fact. For example, Book Plaintiffs' discussion of next-token prediction does not specifically controvert the fact as stated. While next-token prediction is a loss function used in training, that technical "objective" does not compete with the "fundamental objective" of generalization and is in fact the technical mechanism through which model parameters are updated and statistical relationships between tokens are embedded in the model, which allow the model to respond to new, previously unseen inputs. Books Plaintiffs' expert Zhao describes next-token prediction as "[t]he optimization objective." ECF 1787 (Zhao Decl.) Ex. B, ¶ 26. In the same paragraph he concedes that "the model will exhibit some amount of generalization" and that patterns in the training data encoded in the weights enable next-token prediction. *Id*. Dr. Zhao also agrees that in training "[u]ltimately, the model will learn the most common patterns in the grammar, syntax, and semantics of a given language." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 29:15-24. This aligns with Microsoft's LLM expert, Dr. Lafferty's, opinion that "next-token prediction provides a self-supervised learning signal that allows the model to learn grammar, syntax, factual associations, style, and other statistical regularities across the corpus. The purpose of training an LLM is to learn patterns that generalize beyond the text examples shown during training. Through training on large and diverse collections of text, an LLM learns statistical relationships that can be applied to new prompts and

Fourth, **admissible** evidence proves that Defendants trained the at-issue LLMs not for an "unlimited range of contexts and tasks", but rather to become capable of book-like outputs. Zhao Decl. Ex. B ¶ 137.	contexts." ECF 1723 (Lafferty Decl.) ¶¶ 16-17, 20-21, 272. And "[l]ikewise, in language modeling, the goal is not to always perfectly predict the next word or token in the training data, but to accurately model the probability of the next word in new, previously unseen contexts." *Id.*, ¶ 276. Books Plaintiffs' cited evidence does not refute that the fundamental objective of LLM training is generalization. The quoted language that Plaintiffs attribute to Mr. Tworek comes from an internal OpenAI message thread also quoted in BP SUMF ¶ 560. For the reasons stated in Defendants' Response to BP SUMF ¶ 560, that evidence fails to create a dispute that OpenAI trains neural networks with the objective of memorizing datasets. Likewise, Plaintiffs' assertions that "to the extent LLMs generalize at all they do so by encoding the expression contained in their training data" and that generalization is merely "hoped-for" are improper legal argument. Finally, Books Plaintiffs' cited evidence does not controvert that the LLMs at issue were trained for an "unlimited range of contexts and tasks." Dr. Zhao's paragraph 137 consists of hearsay selectively quoting one OpenAI employee's statements about *A Song of Ice and Fire*. That does not show that the models were trained for the purpose of producing book-like outputs rather than to be a general-purpose tool for an "unlimited range of contexts and tasks."

Paragraph 10

10. The capacity to generalize is an emergent property of learning across the complete distribution of the training data. No individual work is responsible for conferring that capability. Lafferty Decl. ¶¶ 16, 33-34, 45, 66, 80-81, 99, 145-146; *see also* Hurst Decl. Ex. 81 (Nadella Dep. Tr.) 90:9-23, 274:12-275:12.	
Class Plaintiffs' Response: Disputed. Microsoft identifies no evidence that the capacity to generalize is an "emergent property" of learning across "the complete distribution of the training data" rather than a product of encoding the individual works within it, and the record is to the contrary: what a model retains scales with its capacity rather than converging on structure shared across the corpus. Researchers found that "larger language models consistently memorized more training data than smaller LMs," with the 1.5 billion parameter GPT-2 model memorizing "over 18 x as much content as the 124 million parameter model." Nelson Decl. Ex. 950 at 2645. OpenAI's own expert likewise acknowledges that "language models memorize up to a certain capacity." Nelson Decl. Ex. 951 at ¶ 63. **Further Disputed** insofar as the statement asserts that generalization is the objective of model training. As set forth in greater detail in response to Microsoft's Statement of Fact No. 9, the training objective is to match the training data as closely as possible. **Further Disputed** that individual works do not contribute to a model's capacity: each individual book used as training data leaves its own imprint on the LLM and contributes to its ability. As Dr. Shan explained, "memorization and generalization are not opposites," "memorizing training examples is mathematically necessary for near-optimal generalization," and "[t]he capacity to memorize also grows with model scale." Shan	**Microsoft's Reply:** Paragraph 10 is not genuinely disputed. **Undisputed** "that no single work, standing alone, is responsible for a model's capacity to generate text." Plaintiffs fail to create a dispute of fact that an LLM's capacity to generalize is "a product of encoding individual works" rather than "an emergent property of learning across the complete distribution of the training data." *First*, Plaintiffs' cited evidence does not specifically controvert the fact as stated, and Plaintiffs mischaracterize it. Plaintiffs are using the word "capacity" to mean size when the fact as stated plainly uses the word "capacity" to mean capability—that the capability to generalize comes from the acquisition of patterns from the training data, and patterns are functions of data distribution not any individual work. Plaintiffs quote Dr. Callison-Burch as acknowledging that "language models memorize up to a certain capacity," but the full sentence states that "while 'language models memorize up to a certain capacity,' further training causes models to 'substitute unintended memorization for generalization' and to 'learn general, reusable patterns as opposed to sample-level specifics." Nelson Decl. Ex. 951 (C. Callison-Burch Reply Expert Rpt.) ¶ 63. The Carlini paper attacked GPT-2 (not the models at issue) to study memorization; it does not address the source of a model's capacity to generalize, and the cited page reports that "among 600,000 (honestly) generated samples, our attacks

Decl. Ex. D ¶ 38. **Undisputed** only that no single work, standing alone, is responsible for a model's capacity to generate text.	find that at least 604 (or 0.1%) contain memorized text." Nelson Decl. Ex. 950 (Extracting Training Data from Large Language Models Research Paper) at -645. Dr. Shan's paragraph 38 relies on "Feldman 2020's analysis," (ECF 1814 (Shan Decl.) Ex. D ¶ 38), which does not describe or study LLMs, but rather, studies an image classification algorithm and memorization of rare, atypical training images in small image-classification datasets—not language modeling (ECF 1723 (Lafferty Decl.) ¶¶ 416-17, 420-22). Further, Dr. Shan testified that he does not "opine on the LLMs generalized in my report" and performed no analysis of model parameters or weights. Hurst Decl. Ex 352 (S. Shan Tr.) at 313:6-13, 315:6-14. Dr. Zhao likewise testified that "there is no easy mapping between specific model weights or parameters and certain concepts or features." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 354:11-16. None of this testimony controverts the fact that generalization arises from patterns across the distribution of data. *Second*, Plaintiffs' assertion that "each individual book … leaves its own imprint on the LLM" is unsupported legal argument and is unsupported by the evidence as described above. *Third*, the dispute about the objective of training is addressed in Microsoft's Reply regarding Paragraph 9.

Paragraph 11

11. Because the statistical patterns a model learns are distributed redundantly across millions of sources spanning trillions of tokens, the removal of any single work or collection of works from the training corpus would not materially alter the statistical structure the model learns. Lafferty Decl. ¶ 33-34, 99, 322.

Class Plaintiffs' Response: Disputed to the extent Microsoft states that removing a category of works from the training corpus would not materially alter the statistical structures the model encodes with respect to that category. OpenAI directly tested whether public domain books could substitute for the LibGen books in its training corpus, and OpenAI employee Qiming Yuan reported the result: "It doesn't seem like public domain books can replace lib[gen] without losing performance. By replacing libgen 1 and 2 with public domain books we are seeing about 3% drop on [redacted]." Nelson Decl. Ex. 289 at -367. As Dr. Zhao explains, Yuan found both a drop on a benchmark of writing quality and a compute penalty—that is, the model trained less efficiently even using the same quantity of processing power—as a result of the swap. Zhao Decl. Ex. A ¶¶ 27-29; *see* Nelson Decl. Ex. 289 at -367–371 (noting the compute penalty cost from the removal and significant degradation in certain evaluations); Nelson Decl. Ex. 952 at 152:25–154:19 (discussing an evaluation he conducted that found decreased model performance when public domain books replaced LibGen books in training data); Nelson Decl. Ex. 67 at 149:25–155:24 (discussing Nelson Decl. Ex. 289 at -367–371); Nelson Decl. Ex. 280 (January 2022 exchange in which Greg Brockman sought the results of removing books from the training set on training efficiency metrics and Nick Ryder responded that the removal caused a 20-40% hit on [redacted]). That result is consistent with the distinct role books play in model capability. Books are crucial sources of long-form text that enable LLMs to process coherent long-form text as an input and produce it as an output, and they provide expressively varied text that enables LLMs to produce text across a wide array of genres, subjects, and styles. Zhao	**Microsoft's Reply:** Paragraph 11 is not genuinely disputed. **Undisputed** that the statistical patterns a model learns are distributed redundantly across millions of sources spanning trillions of tokens. Nor do Books Plaintiffs dispute that the removal of any single work from the training corpus would not materially alter the statistical structures the model learns; they limit their dispute to the removal of "a category of works" and its hypothetical effect "with respect to that category." Plaintiffs do not contradict the fact as stated. The fact states that the removal of any single work or ***collection of works*** (in the copyright law sense of "a work … in which a number of contributions, constituting separate and independent works in themselves, are assembled into a collective whole," 17 U.S.C. § 101) would not materially alter the statistical structure the model learns. It says nothing about the removal of an entire ***category*** of works, such as all books. Plaintiffs assert that removing the category of books materially alters the model, which does not create a genuine factual dispute because it is unsupported by the cited evidence and contradicts Plaintiffs' own expert. Dr. Zhao testified that the evidence from Mr. Yuan on which Plaintiffs rely stated that substituting books "registered a drop in a performance on a benchmark but that, you know, speaking high level that model was still trained and it was still functional." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 326:2-18. He also testified that "you do not need to train on books explicitly for any LLM to be functional." *Id.* at 169:14-21. A change on one benchmark is not evidence that the statistical structure the model learns was materially altered. Similarly, Mr. Brockman's testimony that training on data

Decl. Ex. A ¶ 12(a)(iv); *see also id.* ¶ 21 (OpenAI sought books from Microsoft because books improve an LLM's ability to process "[l]ong term dependencies"); *id.* ¶ 64 (machine-unlearning research identifying "books as high-impact data with disproportionate influence on model capabilities relative to their share of training tokens"). OpenAI's President Greg Brockman confirmed the general principle that dataset composition drives capability: "[w]hen you train these models on data for – they're relevant to specific areas, broad areas, like coding, they get better at that area." Nelson Decl. Ex. 953 at 300:12–18; *see also* Nelson Decl. Ex. 954 (OpenAI internal document stating that "[w]hen thinking about whether or not to include a new dataset in our pre-training runs, there should be only two considerations," including whether "the capability that it advances [is] something we want").	relevant to "broad areas, like coding" improves performance in "that area," Nelson Decl. Ex. 953 (G. Brockman Tr.) at 300:15-18, concerns adding data from a broad domain, not removing any work or collection of works. Further, his testimony supports the idea that including a category of training data—like computer code—in the training corpus improves the model's ability to perform in that area (coding). It does not support that removal of a collection of computer code—say code from one programmer—from the training corpus would materially alter the statistical structure of the model.

Paragraph 12

12. A properly trained LLM is a set of billions or even trillions of numerical parameters (mathematical weights) that encode generalized statistical relationships of language such that the model can simulate conversation and reasoning. Lafferty Decl. ¶¶ 16, 33-35, 45, 52, 71-72, 74-76, 90-91, 94-100, 139-40, 154-55, 157-58, 164; *see also* Hurst Decl. Ex. 81 (Nadella Dep. Tr.) 90:9-23, 275:4-276:13; Ex. 103 (Russinovich Dep. Tr.) 106:4-107:15, 9:3-14, 9:18-23, 12:24-13:15, 14:18-15:17, 16:5-10; Ex. 16 (Brockman Dep. Tr.) 319:20-320:12, 23:19-23, 24:3-8, 26:22-27:9, 37:9-38:8; Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 50:10-51:23, 181:25-182:4; Ex. 117 (Sutskever Dep. Tr.) 27:4-15, 28:5-29:15, 33:17-34:9, 67:9-15, 255:12-18; Ex. 96 (Radford Dep. Tr.) 20:25-23:3.	
Class Plaintiffs' Response: Disputed insofar as the statement implies there is only one way to "properly train" an LLM. **Undisputed** that an LLM can be "a set of billions or even trillions of numerical parameters (mathematical weights) that encode generalized statistical relationships of	**Microsoft's Reply:** Paragraph 12 is not genuinely disputed. **Undisputed** "that an LLM can be 'a set of billions or even trillions of numerical parameters (mathematical weights) that encode generalized statistical

language such that the model can simulate conversation[.]" **Further Disputed** as to "conversation" and "reasoning," as set forth in greater detail in response to Microsoft's Statement of Fact No. 3. Indeed, Microsoft published its own evaluation of nine widely used large language models and found, among other things, that "the ability of these models to output factual text is related to their ability to memorize it" and "higher-quality discourse is usually returned by the models that memorize the most." Nelson Decl. Ex. 956 at 3. Microsoft's own word—that "higher-quality discourse" is effectively a function of a model's capacity for memorization—confirms that what the models produce is an imitation of reasoning rather than reasoning itself.	relationships of language such that the model can simulate conversation[.]"" Plaintiffs' again attempt to create a dispute of material fact over the use of the term "reasoning," but as Microsoft explained in its Reply to Paragraph 3, Microsoft identifies reasoning in the empirical sense of describing the natural language processing tasks LLMs can perform; it makes no claim about whether LLMs "reason" in accordance with human psychology. Indeed, Plaintiffs' own characterization, that the models produce "an imitation of reasoning," is consistent with, not contrary to, Paragraph 12's statement that a model can "simulate" reasoning. There is no genuine dispute. There is also no material factual dispute regarding whether there is "only one way to 'properly train' an LLM." Paragraph 12 does not state that there is only one way to train an LLM. Nelson Decl. Ex. 956 also does not create a dispute of material fact. The paper states that memorization in LLMs may be "simply a training issue," for instance if data isn't deduplicated. Nelson Decl. Ex. 956 (An Evaluation on Large Language Model Outputs: Discourse and Memorization Microsoft Research Paper) at 12. This is consistent with Dr. Lafferty's opinion that properly trained models generalize and that memorization is an unwanted and unintended consequence that occurs when a model is trained on the same data too many times and thus overfits to the training data. ECF 1723 (Lafferty Decl.) ¶ 157. Further, the paper never says outputs are an imitation of reasoning, or that quality is "a function of" memorization. That is improper attorney argument added to the exhibit.

	Furthermore, Plaintiffs overstate the findings of the paper which suggests that for some models (many of which are not at issue) the ability to output factual text is related to their ability to memorize it. Nelson Decl. Ex. 956 (An Evaluation on Large Language Model Outputs: Discourse and Memorization Microsoft Research Paper) at 3. Similarly, Plaintiffs' assertion of quality being "effectively a function of" memorization overstates a tendency on which the paper hedges. *See id.*

Paragraph 13

13. When generalization is strong, an LLM generates coherent and contextually appropriate text about unfamiliar subjects because its parameters represent the deeper structure of language rather than the surface content of the training set; the resulting output is novel text distinct from the source material used for training. Lafferty Decl. ¶¶ 76-77, 81, 141, 145, 155-56; *see also* Hurst Decl. Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 48:20-50:19, 104:4-17, 180:21-182:4, 184:18-185:18; Ex. 16 (Brockman Dep. Tr.) 265:19-266:5.	
Class Plaintiffs' Response: Disputed to the extent Microsoft suggests that when an LLM's generalization is "strong" its output will be novel and will not reproduce or regurgitate the data on which it was trained. To the contrary, optimal generalization depends on the model memorizing its training data. Shan Decl. Ex. D ¶ 14 ("models must memorize long-tail training examples to approach optimal performance on naturally-occurring data distributions"); *id.* ¶ 38 ("memorization and generalization are not opposites," and "memorizing training examples is mathematically necessary for near-optimal generalization"); Zhao Decl. Ex. B ¶ 41 (the view that "models do not memorize but rather 'learn' and 'generalize' as humans do is fundamentally inconsistent with the	**Microsoft's Reply:** Paragraph 13 is not genuinely disputed. **Undisputed** that "[w]hen generalization is strong, an LLM generates coherent and contextually appropriate text about unfamiliar subjects because its parameters represent the deeper structure of language rather than the surface content of the training set." Books Plaintiffs' experts concede that a model "will exhibit some amount of generalization." ECF 1787 (Zhao Decl.) Ex. B ¶ 26; Hurst Decl. Ex. 370 (B. Zhao Tr.) at 338:3-13. Books Plaintiffs try to create a dispute as to whether "when generalization is strong … the resulting output is novel text distinct from the source material used for training," but their assertions do not specifically controvert the stated fact and are not supported by their cited evidence or the record.

mathematical process at the heart of LLM training: lossy compression"). Microsoft's own technical expert has described language models as doing the opposite of what this statement asserts: Dr. John Lafferty wrote that language models "predict[] or, equivalently, compress[] natural text," Nelson Decl. Ex. 899 at ix. Ilya Sutskever, one of OpenAI's co-founders and its former Chief Science Officer, made the same point in research memoranda sent to Bill Gates as part of OpenAI's efforts to induce Microsoft to invest: the "training of GPT-2-type models is compression," a "GPT-2-type model is maximizing the likelihood of all its training data, which happens to be almost precisely equivalent to compressing it," and "compressing the training data with large and deep neural networks leads to 'optimal unsupervised learning.'" Nelson Decl. Ex. 901 at -111, -134–135. That remains OpenAI's position: asked publicly by Sam Altman "why does unsupervised learning work," OpenAI research lead Daniel Selsam answered "compression," explaining that "what we're doing with pretraining . . . is compressing, trying to find the shortest program that explains all of the data that humans have produced." Zhao Decl. Ex. B ¶ 33. Second, it is **Disputed** that an LLM's "resulting output is novel text distinct from the source material used for training," as set forth in greater detail in response to Microsoft's Statement of Fact No. 14.	Plaintiffs' assertion that "optimal generalization depends on the model memorizing its training data" does not controvert the stated fact. Plaintiffs' assertion goes only to what aspects of the LLM training process *cause* "optimal generalization," while the stated fact concerns the expected *output* of a fully-trained LLM in which "generalization is strong." The mere occurrence of memorization during the training process does not mean that the resulting LLM will "reproduce or regurgitate the data on which it was trained," as Plaintiffs suggest, rather than generate novel text. Plaintiffs' cited evidence does not address whether an LLM in which "generalization is strong" will generate novel text. Shan Decl. D ¶¶ 14 & 38 say nothing about a model's ability to generate novel text, and the paper on which the quoted portions of both paragraphs rely (Vitaly Feldman, "Does Learning Require Memorization? A Short Tale about a Long Tail") does not describe or study LLMs, but rather, studies an image classification algorithm and memorization of rare, atypical training images in small image-classification datasets—not language modeling (ECF 1723 (Lafferty Decl.) ¶¶ 416-17, 420-22). Defendants moved to exclude Zhao's "lossy compression" memorization opinion contained in Zhao Decl. Ex. B ¶ 41 pursuant to Federal Rule of Evidence 702, and so to the extent the Court excludes the challenged opinion, Plaintiffs may not rely on it to dispute this fact. ECF 2050; ECF 2055 at 3. Further, whether and to what degree "compression" occurs during the training process and its relationship to generalization does not controvert the stated fact, which concerns the expected output of an LLM in which "generalization is strong." Plaintiffs' cited evidence does not establish that the alleged occurrence of compression

	during the training process is inconsistent with a finished model in which "generalization is strong" being able to output novel text. Nelson Decl. Ex. 899 is a preface to a compendium collecting various technical papers on machine learning from 2003—well before LLMs existed. *See* Nelson Decl. Ex. 899 (Lafferty Dep. Ex. 8 – Language Modeling for Information Retrieval Paper) at iv. Pursuant to FRE 106, the preface to the compendium co-authored by Dr. Lafferty states "the first statistical language modeler, Claude Shannon, "[i]n exploring the application of his newly founded theory of information to human language…measured how well simple *n*-gram models predicted or, equivalently, compressed natural text," citing to Mr. Shannon's work published *in 1951*. *Id.* at ix. At his deposition, Dr. Lafferty explained that this sentence does not imply that language *modeling* is equivalent to compression, and that while a language model could be used as a compression algorithm to compress data, one could not go in the other direction and decompress training data out of an LLM. Hurst Decl. Ex. 374 (J. Lafferty Tr.) at 201:5-204:22, 206:21-211:19 (emphasis added). The cited statement by Dr. Lafferty, as well as his testimony--while asserting that a properly trained LLM is not a compressed form of its training data—is not inconsistent with the possibility of compression occurring during the training process. Similarly, the other evidence cited by Plaintiffs (Nelson Decl. Ex. 901 and Zhao Decl. Ex. B ¶ 33) discuss the possible role of "compression" in "pretraining." This evidence alleging compression occurs during the training process does not contradict that a finished model in which "generalization is strong" will be able to output novel text.

Paragraph 14

14. Models can at times generate text that was part of their training corpus, a known phenomenon called regurgitation. However, instances of regurgitation in a properly trained model are extremely rare. A model that has a tendency to regurgitate its training data reflects an unintended side-effect of the training process rather than the intended operation of the model. Lafferty Decl. ¶ 35; *see also* Usdan Decl. ¶¶ 19-20.

Class Plaintiffs' Response: Undisputed that models "can at times generate text that was part of their training corpus, a known phenomenon called regurgitation." **Disputed** that "instances of regurgitation in a properly trained model are extremely rare" and that regurgitation "reflects an unintended side-effect of the training process rather than the intended operation of the model." At the threshold, the models at issue were not "properly trained" within the meaning of Microsoft's assertion: OpenAI released GPT-3 and GPT-3.5 without anti-regurgitation protections at all. Nelson Decl. Ex. 957 at -166 ("GPT 3.5 was not trained with regurgitation fixes"). And on March 13, 2023, OpenAI employee Peter Welinder flagged "refusals to avoid regurgitated copyright text" as a "fairly high priority update[]" that OpenAI "need[ed] for the GPT-4 model"; OpenAI released GPT-4 the following day without implementing it. Nelson Decl. Ex. 904 at -363; Shan Decl. Ex. D ¶ 47. OpenAI's own evaluation of its output filter reported that the filter caught only 4–30% of cases on OpenAI's internal book test sets, meaning that 70–96% of the instances in which the model reproduced copyrighted training text passed through undetected. Shan Decl. Ex. D ¶ 47. Microsoft's own internal benchmarking confirms that regurgitation is neither rare nor incidental. *See* Nelson Decl. Ex. 902 at -121 [redacted]	**Microsoft's Reply:** Paragraph 14 is not genuinely disputed. **Undisputed** that models "can at times generate text that was part of their training corpus, a known phenomenon called regurgitation." Plaintiffs' evidence (a) does not controvert that regurgitation in a properly trained model is extremely rare and is an unintended side effect of training, (b) is inadmissible, and/or (c) contradicted by their own expert's testing. Plaintiffs' evidence confirms that the rate of regurgitation is rare. Dr. Shan prompted GPT models adversarially with passages from Plaintiffs' books. Even with manipulative prompting, he produced alleged regurgitation only 0.53% of the time (28,236 of 5,301,835 attempts). Hurst Decl. Ex. 107 (S. Shan Tr.) at 146:8-17. His long-form book testing produced just 28 alleged regurgitations in 13,353 attempts. Hurst Decl. Ex. 352 (S. Shan Tr.) at 184:13-24. Across 8.2 million Copilot logs, he found only 24 conversations with any alleged regurgitation, about 0.00029%. Hurst Decl. Ex. 241 (Shan Opening Report) ¶ 156. The Carlini paper Plaintiffs cite is to the same effect: it found regurgitated text in "at least 604 (or 0.1%)" of "600,000 (honestly) generated samples." Nelson Decl. Ex. 950 (Extracting Training Data from Large Language Models Research Paper) at 645. By any reasonable measure, these rates are "extremely rare."

[redacted]).

Microsoft further found that outputs reflecting some degree of memorization were "more likely to be considered high quality," that "memori[z]ation and discourse quality [were] very, very correlated," and that OpenAI's models performed as they did on memorization because "OAI is definitely overfitting their models." Nelson Decl. Ex. 184 at -752, -743. As Microsoft's CTO of Azure, Mark Russinovich, has noted: "[I]t's easy to get DV3 to generate copyrighted content. Here I aimed at specific stories but it's clear that it would be easy for someone to stumble onto copyrighted output." Nelson Decl. Ex. 958 at -983. In his deposition, Mr. Russinovich similarly stated that he "probed [the model's] ability to generate verbatim passages from works like Harry Potter," and found that "I was able to get the model to generate verbatim text." Nelson Decl. Ex. 959 at 104:16-105:2. OpenAI likewise acknowledged internally that "models want to regurgitate, so we're inherently fighting against the training objective." Nelson Decl. Ex. 298 at -468.

Class Plaintiffs' expert Dr. Shan further confirmed that verbatim reproduction or "regurgitation" is a "predictable result of the training objective (which drives the model toward reproducing high-likelihood training sequences), rather than a corner-case behavior that arises only under adversarial prompting," and that "models must memorize long-tail training examples to approach optimal performance on naturally-occurring data distributions." Shan Decl. Ex. D ¶¶ 45, 14. He also found that verbatim reproduction "also surfaced during two independent

None of Plaintiffs' other evidence speaks to the rate. Mr. Russinovich's message and testimony describe isolated outputs from a model, not from the Copilot product. Isolated outputs say nothing about how often regurgitation occurs. Nelson Decl. Exs. 958 (MSFT _AI_MDL_002836983), 959 (M. Russinovich Tr.).

Nelson Decl. Ex. 902 (MSFT _AI_MDL_001885086) is also inadmissible, for three reasons:

- *No foundation*. It appears to be an unauthenticated draft presentation. When a different version was shown to Xia Song at his deposition, he did not recall the document and did not know what it was. He was not asked about the cited page. Hurst Decl. Ex. 375 (X. Song Tr.) at 147:7-17, 149:23-150:21.
- *Inadmissible Hearsay*. The cited slide is not "Microsoft's own internal benchmarking." It is titled [redacted] and summarizes a third-party study that is not in the record. Ex. 902 (MSFT _AI_MDL_001885086) at -121.
- *Mischaracterized*. The slide does not show that [redacted] The benchmark [redacted] *Id.*

As to the intended operation of the model, Plaintiffs' evidence does not controvert that regurgitation is not the intended operation of the models. Their own exhibit says the opposite: [redacted] ECF 1707

experiments that were not explicitly designed to elicit it," including a summarization test in which "the prompt requests only a high-level summary," and that prompts drawn from the LibGen-format copies of Plaintiffs' works produced longer verbatim reproductions than prompts drawn from optical-character-recognition scans of the same works, *id.* ¶ 31. As Dr. Shan put it, "[t]he fact that regurgitation appears even when no user attempts to extract it shows that memorization is a structural property of the models themselves, rather than an artifact produced by adversarial prompting." *Id.* ¶ 31. Defendants also had reason to know that their models could and would regurgitate training data: OpenAI's internal records identify the anti-regurgitation team's ultimate backstop as "rely[ing] on the 'client' to do the right thing," and document an OpenAI engineer "disabling anti-regurg for the weekend," to which Greg Brockman replied "let this be the last weekend where that's true!" Shan Decl. Ex. D ¶ 49.

Further, the cited evidence does not support the proposition that regurgitation is an unintended side effect of the training process. [redacted] Usdan Decl. ¶ 19; *id.* ¶ 20 [redacted]

(Usdan Decl.) ¶ 20. Plaintiffs also rely on OpenAI's "anti-regurgitation team" and safeguards. A company has no reason to build either if regurgitation is the model's intended function. Microsoft built similar safeguards "to address the risk of this unintended outcome, whether user-elicited or resulting from a system defect." ECF 1710 (Bird Decl.) ¶ 35.

Plaintiffs also treat "memorization" and "regurgitation" as if they were the same thing. They are not. Memorization during training does not mean regurgitation will follow, and it does not make regurgitation the model's intended operation. The Microsoft research paper Plaintiffs cite addresses memorization only. It found that a model trained on "a properly-deduplicated, evenly-distributed, sufficiently-representative dataset would (in theory) not have a correlation between text uniqueness, memorized content, and discourse quality." Nelson Decl. Ex. 184 (An Evaluation on Large Language Model Outputs: Discourse and Memorization Microsoft Research Paper) at -762.

Dr. Shan concludes that regurgitation without user extraction attempts "shows that memorization is a structural property of the models themselves, rather than an artifact produced by adversarial prompting." ECF 1814 (Shan Decl.) ¶ 31. That conclusion does not contradict Paragraph 14, which does not mention adversarial prompting. It is also undercut by his own testing, which produced almost no regurgitation.

Plaintiffs argue that particular models were not "properly trained." This does not controvert the fact. At most, Plaintiffs' evidence shows which anti-regurgitation measures were used in which models. It actually supports

[redacted]	the fact: regurgitation is an unintended side effect that additional training or other mitigations is designed to correct, not the intended operation of a properly trained model. And Dr. Shan's testing of the models at issue found regurgitation only in extremely rare cases, as shown above. Plaintiffs' remaining assertions about OpenAI's release timing and internal mitigation work are an improper recitation of facts. They do not specifically controvert Paragraph 14 and are unsupported by the record, for the reasons set out in Defs.' Resp. to BP SUMF ¶ 564.

Paragraph 15

15. Books Plaintiffs' expert Dr. Shan analyzed logs produced by Microsoft of 8.2 million real-world conversations with Microsoft's Copilot in order to assess whether any of those Copilot outputs contained what he qualified as a "regurgitation" of Books Plaintiffs' Asserted Works. Shan's definition of a "regurgitation" was a contiguous near-verbatim block of at least 30 consecutive words (with certain other exclusions):	
Class Plaintiffs' Response: Disputed to the extent Microsoft attempts to provide an exhaustive account of Dr. Shan's methodology, including his definition of "regurgitation." Shan Decl. Ex. D § IV. **Further Disputed** to the extent Microsoft suggests that the 8.2 million conversations constitute a representative or complete sample. The logs were selected and produced by Microsoft, and Dr. Shan's findings within that sample are a lower bound on the models' reproduction of the Asserted Works, as set forth in greater detail in response to Microsoft's Statement of Fact No. 15(d). Otherwise, **Undisputed**.	**Microsoft's Reply:** Paragraph 15 is not genuinely disputed. **Undisputed** that Dr. Shan "analyzed logs produced by Microsoft of 8.2 million real-world conversations with Microsoft's Copilot" to assess whether those outputs contained a "regurgitation" of the Asserted Works, or that his definition of "regurgitation" was "a contiguous near-verbatim block of at least 30 consecutive words." However, Plaintiffs seek to create a dispute of fact that Paragraph 15 gives an incomplete account of Dr. Shan's methodology and that the 8.2 million conversations are not representative. The cited evidence does not specifically controvert the fact as stated. Paragraph 15 does not purport

	to give an exhaustive account—it expressly notes "certain other exclusions"—and it says nothing about whether the sample is representative. Section IV of Dr. Shan's corrected rebuttal report, ECF 1814 (Shan Decl.) Ex. D § IV, cited by Class Plaintiffs here, concerns his testing of OpenAI models; his Copilot log methodology appears in his opening report, which confirms the 30-word reporting threshold. ECF 1814 (Shan Decl.) Ex. A ¶¶ 38-40, 155-156.
a. For 202 of the 212 Asserted Works, Dr. Shan identified no "regurgitations" in the 8.2 million Copilot logs;	
Class Plaintiffs' Response: Disputed. Admissible evidence contradicts Microsoft's statement. Microsoft's log production is a sample of Copilot conversations selected by Microsoft, and the absence of a matching output within that sample does not establish that a model is incapable of reproducing a given Asserted Work. Shan Decl. Ex. F ¶ 56 ("Because I understand that the produced logs were not assembled using search terms assembled by the Class Plaintiffs in this case, the set which I searched may itself under-represent regurgitation."). **Undisputed** only that output log samples produced by Microsoft in this case include regurgitations of at least 10 of Plaintiffs' Works, which is a floor rather than a measure of the models' reproduction capacity. Dr. Shan's testing of the models themselves identified 437 distinct verbatim reproductions of Plaintiff-book text drawn from 43 of the 212 Plaintiff books and all 14 Plaintiff authors, and Dr. Shan explained that his log-based search can only understate the true rate, as set forth in greater detail in response to Microsoft's Statement of Fact No. 15(d). Shan Decl. Ex. D ¶ 22; Shan Decl. Ex. F ¶ 56.	**Microsoft's Reply:** Paragraph 15(a) is not genuinely disputed. **Undisputed** that the produced logs "include regurgitations of at least 10 of Plaintiffs' Works." That concession is the mathematical inverse of Paragraph 15(a)'s factual assertion that "[f]or 202 of the 212 Asserted Works, Dr. Shan identified no 'regurgitations' in the 8.2 million Copilot logs." The rest of Plaintiffs' response does not controvert the stated fact and is improper legal argument going to the weight that should be given to the stated fact. Paragraph 15(a) reports what Dr. Shan found in the logs. It makes no assertion about how representative the log production is or what the models are capable of reproducing. The response nonetheless argues three things: that the logs are a sample "selected by Microsoft," that the absence of a match "does not establish that a model is incapable of reproducing" a work, and that the logs "may" under-represent regurgitation. Each of these are improper legal arguments that answer a proposition Microsoft did not state. Similarly, Dr. Shan's separate testing of the OpenAI models

	is a different analysis of different data. It does not controvert his findings from the Copilot logs. Lastly, Class Plaintiffs' response to Paragraph 15(d), to which Class Plaintiffs cross-reference, suffers from the same flaws, and so, too, is an improper response.
b. **Of the remaining 10 works, out of the 8.2 million conversations, Dr. Shan identified a total of 24 conversations meeting his definition of containing one or more "regurgitations," a rate of .00029%;**	
Class Plaintiffs' Response: Undisputed that output log samples produced by Microsoft in this case include at least 24 instances of conversations containing regurgitations. **Disputed** that a regurgitation rate can be extrapolated from that sample. **Further Disputed** that .00029% measures the models' propensity to reproduce the Asserted Works, as set forth in greater detail in response to Microsoft's Statement of Fact No. 15(d).	**Microsoft's Reply:** Paragraph 15(b) is not genuinely disputed. **Undisputed** that out of the 8.2 million conversations, Dr. Shan identified a total of 24 conversations meeting his definition of containing one or more "regurgitations." However, Plaintiffs seek to create a dispute of fact that "a regurgitation rate can[not] be extrapolated from that sample" and that .00029% does not measure the models' propensity to reproduce the Asserted Works. That purported dispute does not create a genuine issue of material fact. Paragraph 15(b) does not extrapolate; it states the arithmetic result of Dr. Shan's own findings—24 conversations out of 8.2 million, or approximately .00029%.
c. **To arrive at those 24 conversations, Dr. Shan's search code combined the "Text" field, which contains Copilot responses displayed to users, and the "HiddenText" field, which contains web-grounding material not displayed to users; the search did not distinguish between the two;**	
Class Plaintiffs' Response: Disputed to the extent Microsoft purports to provide an exhaustive account of Dr.	**Microsoft's Reply:** Paragraph 15(c) is not genuinely disputed. **Undisputed** that, to arrive at the 24

Shan's methodology. Shan Decl. Ex. D § IV. **Further Disputed** insofar as Microsoft contends that "web-grounding material not displayed to users" does not represent an instance of the model or inference process regurgitating Class Plaintiffs' Works. *See* Shan Decl. Ex. F ¶ 59. Otherwise, **Undisputed**.	conversations, Dr. Shan's search code combined the "Text" and "HiddenText" fields and did not distinguish between them. Plaintiffs also do not dispute that the "HiddenText" field "contains web-grounding material" or that such material is "not displayed to users." Plaintiffs seek to create a dispute of fact only that (1) this fact is not an exhaustive account of Dr. Shan's methodology and (2) web-grounding material not displayed to users represents an instance of the model regurgitating the Asserted Works. As to the first point, there is no dispute because the fact does not purport to be an exhaustive account of Dr. Shan's methodology. As to the second point, Plaintiffs' assertion that the undisplayed web grounding material reflects model regurgitation does not controvert the stated fact, is not supported by the cited evidence, and is contradicted by Plaintiffs' own expert. Dr. Shan nowhere identifies evidence that the "HiddenText" material was generated by a model. Nor could he, as grounding data comes from internet retrieval and so cannot be regurgitation from an LLM. Dr. Shan's own opening report describes grounding material as "internal search and retrieval logs, which also reflect crawling done in the course of answering queries." ECF 1814 (Shan Decl.) Ex. A ¶ 155.
d. **Shan found no matches of entire books or book chapters when performing his analysis.** Lafferty Decl. ¶¶ 366-68, 378; Hurst Decl. Ex. 241 (Shan Opening Report) ¶¶ 40, 156 (24 responses; 10 works; 30-or-more near-verbatim words); Hurst Decl. Ex. 243 (Shan Reply Report) ¶ 59.	
Class Plaintiffs' Response: Disputed to the extent Microsoft extrapolates from the sample or uses this statement as a measure of its models' capacity to reproduce	**Microsoft's Reply:** Paragraph 15(d) is not genuinely disputed. **Undisputed** that Dr. Shan "found no matches of entire books or book chapters when performing his

the Asserted Works. As Dr. Shan explained: "The search misses genuine reproductions for at least three independent reasons: which conversations I could examine, what I compared them against, and how the logs were sampled. Because of the format of the production and computational limits, I searched only the conversations that named an Asserted Work's author and set the rest aside, so any reproduction in a conversation that did not name the author was never counted. I also matched against OCR-derived copies of the books, whose scanning artifacts can break exact word matches and cause real reproductions to be missed. Because I understand that the produced logs were not assembled using search terms assembled by the Class Plaintiffs in this case, the set which I searched may itself under-represent regurgitation. Each of these can only lower the count, so the reported rate is a lower bound on how often these models reproduce the books." Shan Decl. Ex. F ¶ 56. That limitation is inherent in log-based searching rather than particular to Dr. Shan's methodology: researchers studying extraction have found that "[m]uch of the training data that we extract is only discovered when prompting the LM with a particular prefix," so that absent the right prompt, memorized content goes undetected. Nelson Decl. Ex. 950 at 2645. Dr. Shan's testing of the models themselves, as distinct from the produced logs, identified "437 distinct verbatim reproductions of Plaintiff-book text," spanning "more than 44,098 reproduced words in aggregate, drawn from 43 of the 212 Plaintiff books and all 14 Plaintiff authors," with reproductions from 11 of those books exceeding 100 contiguous words. Shan Decl. Ex. D ¶ 22. The longest contiguous passage Dr. Shan extracted was a 1,899-word excerpt from *A Game of Thrones*, which "was	analysis." Dr. Shan confirmed that none of the 24 conversations contains a reproduction of an entire book or a complete chapter. Hurst Decl. Ex. 352 (S. Shan Tr.) at 287:2-19. However, Plaintiffs seek to create a dispute of fact that Dr. Shan's log results understate the models' capacity to reproduce the Asserted Works. That purported dispute does not create a genuine issue of material fact. *First*, the cited evidence does not specifically controvert the fact as stated. Paragraph 15(d) reports what Dr. Shan found; it does not extrapolate from the sample. The limitations Dr. Shan describes in his reply report arise from choices he made in designing his own search, and his statement that the logs "may" "under-represent regurgitation" is speculation. ECF 1814 (Shan Decl.) Ex. F ¶ 56. The Carlini paper describes extraction research on GPT-2, not the Copilot logs. Nelson Decl. Ex. 950 (Extracting Training Data from Large Language Models Research Paper) at -645. Dr. Shan's paragraph 22 concerns his testing of OpenAI models, not Copilot. Hurst Decl. Ex. 352 (S. Shan Tr.) at 316:7-17. Even there, the longest passage he reports was [redacted] words, not an entire book or chapter, and it was produced by a prompt instructing the model, [redacted] ECF 1814 (Shan Decl.) Ex. B ¶ 25. His 41% figure was obtained using the GPT-4-base model. Hurst Decl. Ex. 107 (S. Shan Tr.) at 147:18-24. *Second*, Plaintiffs' characterizations of a "floor" or "lower bound" are legal argument, not evidence that controverts Paragraph 15(d).

reproduced in thousands of independent extraction attempts, ruling out the possibility that the match is a coincidence," and Dr. Shan measured a peak extraction rate of 41% for that work. *Id.* ¶¶ 6, 22. Otherwise, **Undisputed**.	

Paragraph 16

16. At OpenAI's request, Microsoft provided OpenAI with portions of the Bing Index, the web index that powers Microsoft's Bing search engine. Hurst Decl. Ex. 118 (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 304:9-16; Ex. 119 (Templeton 30(b)(6) Dep. Tr. (Vol. 2, dated Jan. 21, 2026)) 390:22-391:1; Ex. 38 (Gaggara 30(b)(6) Dep. (dated Aug. 20, 2025)) 44:13-15; Ex. 37 (Gaffney 30(b)(6) Dep. Tr.) 131:20-133:2, 9:2-5, 9:15-10:5, 10:19-25; Ex. 37.A (Gaffney Dep. Ex. 16) 2-3.	
Class Plaintiffs' Response: Undisputed that Microsoft provided OpenAI with the Bing Index as part of what OpenAI's Chief Research Officer Bob McGrew described as a "trade." Nelson Decl. Ex. 779 (explaining that "the idea here [was] that we're giving data to them and they are giving data to us"). In exchange for the Bing Index, OpenAI agreed to pay Microsoft [redacted]. Nelson Decl. Ex. 874 at 97:23–98:24. Microsoft also received from OpenAI "a portion [redacted] Nelson Decl. Ex. 291. **Disputed** to the extent Microsoft characterizes what it transferred as mere "portions" of the Bing Index. Microsoft provided OpenAI with the full offline version of the Bing Index. Nelson Decl. Ex. 875 at 68:25–69:9 ("the Bing Index that [Microsoft] shared with OpenAI on Azure" was "the full size of the index"). Microsoft 30(b)(6) designee Jamie Huynh likewise testified that the "original Bing Index was the index data that we had shared with OpenAI in its full form." Nelson Decl. Ex. 873 at 167:2–10.	**Microsoft's Reply:** Paragraph 16 is not genuinely disputed. **Undisputed** that "Microsoft provided OpenAI with … Bing Index" data or that the Bing Index is "the web index that powers Microsoft's Bing search engine." Books Plaintiffs fail to create a genuine dispute that Microsoft provided OpenAI "the full offline version of the Bing Index" rather than "portions" of it, because the cited evidence does not specifically controvert the fact as stated. Although Ms. Gaggara testified to the quoted language, (Nelson Decl. Ex. 875 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 68:25-69:9), she further testified that Microsoft "filtered everything, which was no-archive/no-cache and DMCA, and then would have handed it to them – whatever that number would be." *Id.* at 70:13-16. The evidence cited in Paragraph 16 also reflects Ms. Gaggara's testimony that Microsoft shared the Bing Index with OpenAI "*[a]fter filtering it out.*" Hurst Decl. Ex. 38 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 44:13-15 (emphasis added); *see also* Blythe Opp. Decl. Ex. 65 (P. Gaggara 30(b)(6) Tr. (Vol. 1,

	dated Jan. 27, 2026)) at 230:22-231:8 ("We removed anything which was 'noarchive/nocache' from them. We removed anything which was DMCA removed from them."); 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."). Further, Mr. Huynh's deposition testimony (Nelson Decl. Ex. 873) cannot be presented in an admissible form. Specifically, Mr. Huynh's deposition testimony is inadmissible speculation as to the meaning of an OpenAI employee's statement. To the extent Plaintiffs contend that Mr. Huynh's statements were made on behalf of Microsoft, such statements were elicited from Microsoft's corporate designee concerning matters outside the scope of topics for which Microsoft designated this witness, and Plaintiffs have not otherwise established that the witness had personal knowledge sufficient to testify on this matter in his individual capacity. Indeed, in response to questioning concerning the contents of URLs stored in the Bing Index, Mr. Huynh answered, "I'm unsure. I would – I would have to speculate, be speculating." Nelson Decl. Ex. 873 (J. Huynh 30(b)(6) Tr.) at 86:14-18. Plaintiffs' response includes additional assertions that do not address or specifically controvert the stated fact. They therefore do not create a genuine dispute of the fact stated in Paragraph 16 and are irrelevant to the determination of the merits of Microsoft's motion for summary judgment.

Paragraph 17

17. The provision of portions of the Bing Index to OpenAI was a step in the process for the purpose of training LLMs. Hurst Decl. Ex. 107 (Shan Dep. Tr.) 318:11-319:9; Ex. 37 (Gaffney 30(b)(6) Dep. Tr.) 131:20-133:2; Ex. 37.A (Gaffney Ex. 16); Ex. 16 (Brockman Dep. Tr.) 323:11-16.	
Class Plaintiffs' Response: Disputed to the extent Microsoft contends that it distributed "portions" of the Bing Index to OpenAI as opposed to the full version, as set forth in greater detail in response to Microsoft's Statement of Fact No. 16. **Further Disputed** that the provision of the Bing Index to OpenAI was *solely* a step in the process for the purpose of training LLMs. *See* Nelson Decl. Ex. 960 at 66:5–68:25, 158:14–159:7; Nelson Decl. Ex. 867 at 153:13–155:21 (Microsoft and OpenAI both used the Bing Index for retrieval augmented generation); Nelson Decl. Ex. 961 at 131:17-22 (when ChatGPT first began using retrieval augmented generation to respond to user queries, the first source it used was the Bing Index); Nelson Decl. Ex. 962 at slide 4 (Microsoft's Bing Index "powers ChatGPT browsing").	**Microsoft's Reply: Undisputed** that Microsoft's provision of Bing Index data to OpenAI "was a step in the process for the purpose of training LLMs." Books Plaintiffs fail to create a genuine dispute that Microsoft provided OpenAI the "full version" of the Bing Index as opposed to "portions" of it, for the reasons set forth in Microsoft's Reply to Paragraph 16. Books Plaintiffs' assertion that "the provision of the Bing Index to OpenAI" was not "*solely* a step in the process for the purpose of training LLMs" does not controvert the stated fact and is not supported by Plaintiffs' cited evidence. The evidence cited as Nelson Decl. Ex. 960 states that "[redacted] which is a step in the training process. Nelson Decl. Ex. 960 (N. Ryder 30(b)(6) Tr.) at 68:10-25. That deposition testimony does not controvert the evidence cited in Paragraph 17. *See* Hurst Decl. Ex. 107 (S. Shan Tr.) at 319:1-9 ([redacted]). The remaining cited evidence conflates the Bing Index data with the Bing API that ChatGPT used for web grounding. The evidence cited as Nelson Decl. Exs. 867, 961, and 962

	exclusively relate to OpenAI's use of the Bing API to retrieve ordinary Bing search results for grounding ChatGPT answers, and Books Plaintiffs do not contend that the retrieval of web content for grounding is in any way related to the use of content for model training, whether for experimental purposes or developing production LLMs. This evidence bears no relation to Microsoft's provision of portions of the Bing Index to OpenAI for experimentation with training LLMs, which is the subject of the stated fact. Books Plaintiffs have not asserted claims against Microsoft for OpenAI's use of the Bing API for web grounding, nor did they argue or even mention OpenAI's use of the Bing API anywhere in their brief in opposition to Microsoft's motion for summary judgment. These additional facts therefore are irrelevant to the determination of the issues before the Court and do not address or specifically controvert the fact stated. Further, Plaintiffs' characterization that OpenAI's use of the Bing API for web grounding constitutes "provision of the Bing Index to OpenAI" is inaccurate. ChatGPT used the Bing API to retrieve a limited number of Bing search results (including only the title, link, and snippet) most relevant to a user's query to generate web grounded responses; the Bing API is not an access-point to the full contents of webpages contained in the Bing Index. Blanco Opp. Decl. Ex. 161 (E. Abib 30(b)(6) Tr.) at 76:1-25; Blanco Opp. Decl. Ex. 203 (P. Gaggara 30(b)(6) Tr. (Jan. 28, 2026)) at 435:18–437:3; Hurst Decl. Ex. 416 (J. Ribas Tr.) at 235:15-236:6. Microsoft's witness Jordi Ribas testified that Microsoft did not "make a copy of the Bing Index for OpenAI's use in querying [the Bing API]" and explained, "[B]asically, there is the index. Then there is the ranker. And then the Bing API is on top of the ranker and the index, and

	when you send a query to the Bing API, it provides the same results in theory as if you were going to Bing.com." Hurst Decl. Ex. 416 (J. Ribas Tr.) at 236:9-237:24. This characterization is also not supported by the cited evidence. *See* Nelson Decl. Ex. 867 (F. Canel 30(b)(6) Tr.) at 155:3-21 ("leveraging Bing"); Nelson Decl. Ex. 962 at slide 4 ([redacted]). While the deposition testimony cited as Books Ex. 961 reflects that Mr. Perelman agreed with counsel's characterization that ChatGPT used "the Bing Index" for retrieval augmented generation, pursuant to Fed. R. Evid. 106, Mr. Perelman independently testified that [redacted] Hurst Decl. Ex. 376 (A. Perelman 30(b)(6) Tr.) at 30:1-3. Microsoft's corporate designee on topics related to the Bing API services provided to OpenAI testified that characterizing the Bing API as returning content directly from the Bing Index is an inaccurate characterization and explained their respective differences. *See* Hurst Decl. Ex. 377 (E. Abib 30(b)(6). Tr.) at 68:24-71:21.

Paragraph 18

18. Books Plaintiffs' Asserted Works were created to be read, and are sold and licensed to readers for entertainment, information, and edification. Tucker Decl. ¶¶ 11(b), 48, 52-54; Hurst Decl. Ex. 35 (Franzen Dep. Tr.) 21:18-22:18; Ex. 9 (Baldacci Dep. Tr.) 79:7-14; Ex. 128 (Wolicki Dep. Tr.) 21:1-6, 162:16-163:8.	
Class Plaintiffs' Response: Undisputed that the Asserted Works were created and that they are sold and licensed. **Disputed** to the extent Microsoft limits the scope of licenses to "entertainment, information, and	**Microsoft's Reply: Undisputed** that "Books Plaintiffs' Asserted Works were created to be read" and that the "Asserted Works are sold and licensed to readers for entertainment, information, and edification."

edification." Class Plaintiffs who were asked testified that they would consider licensing their works for generative AI training, confirming that the use Defendants made without permission is one for which a market exists and can be licensed. David Baldacci agreed that he would consider licensing his works for "AI training" if someone approached him about it. Nelson Decl. Ex. 920 at 210:19-211:6. Sylvia Day testified that if asked to license her works for training a large language model, she would "absolutely present [the offer] to [her] team and negotiate" and her "primary goal is to monetize [her] work in every way possible." Nelson Decl. Ex. 919 at 157:15-24. John Grisham agreed that he would be "willing to sit down with an artificial intelligence company and license [his] asserted works for training AI models." Nelson Decl. Ex. 925 at 182:16-22. Taylor Branch testified that he would consider an offer to "licens[e] [his] books for use" as AI training data. Nelson Decl. Ex. 921 at 110:4-7. Stacy Schiff testified that she would be "willing to discuss licensing" her works for AI training. Nelson Decl. Ex. 928 at 195:14-196:2. Michael Connelly, Christopher Golden, Andrew Sean Greer, George R.R. Martin, and Jodi Picoult gave comparable testimony. Nelson Decl. Ex. 922 at 156:17-25; Nelson Decl. Ex. 923 at 89:19-90:7; Nelson Decl. Ex. 924 at 112:17-22; Nelson Decl. Ex. 926 at 140:13-141:3, 141:24-142:5, 151:24-152:10, 153:1-3; Nelson Decl. Ex. 927 at 77:7-22; *see also* Lasinski Decl. Ex. B ¶ 238 (analyzing authors' willingness to participate in the licensing market, including an Authors Guild survey finding that 95 percent of writers think they should be compensated for use of their works in developing generative AI).	While Plaintiffs seek to add an additional fact regarding AI licensing to create a dispute, that does not make this fact disputed. Moreover, the evidence cited by Plaintiffs does not contradict the fact as stated, which refers to how books are actually sold and licensed. None of the cited testimony shows that Generative AI training is a current type of license that any of the Books Plaintiffs have actually engaged in. Plaintiffs merely cite testimony from a handful of authors that they would *consider* licensing their works for Generative AI training. This testimony does not create a dispute that the Asserted Works are (currently) "sold and licensed to readers for entertainment, information, and edification."

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

Paragraph 19

19. The LLM training process deconstructs works into tokens, then analyzes the patterns among various pieces of tokenized content in order to translate those patterns into model weights–all to gain access to underlying facts, ideas, linguistic patterns, and statistical relationships. Lafferty Decl. ¶¶ 77, 80-81, 105-114, 137-39, 141, 144-47, 283; *see also* Hurst Decl. Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 17:2-18:7, 18:12-19:4, 19:12-20:18, 48:25-49:24, 50:10-19, 104:15-17, 176:22-177:23; Ex. 101 (Rossett Dep. Tr.) 18:22-23:2, 67:25-68:9.	
Class Plaintiffs' Response: Undisputed that language data is converted into "tokens" and that LLM models encode numerical values in a network of "parameters" or weights. Otherwise, **Disputed.** The purpose of the training process is not to "gain access to underlying facts, ideas, linguistic patterns, and statistical relationships." Rather, the goal is to adjust the parameters of the LLM so that its outputs match the training data as closely as possible. Zhao Decl. Ex. B ¶¶ 14, 25-27. As set forth in greater detail in response to Microsoft's Statement of Fact No. 13, the models are "not a generalized understanding independent of the training data[;] [t]hey are a compressed encoding of the training data," and "[w]hen the model generates output that resembles its training data in style, structure, or directly reproduces content, it is decompressing data encoded during training." Zhao Decl. Ex. B ¶¶ 26-27. Moreover, the purpose of training includes tuning the LLM's parameters to encode the specific structure of different types of text, including the precise expression used in those texts and the expressive styles of different writers. Zhao Decl. Ex. A ¶ 12(a)(iv). Further, the LLM training process results in verbatim memorization. Shan Decl. Ex. D ¶¶ 22, 31, 45.	**Microsoft's Reply:** Paragraph 19 is not genuinely disputed. **Undisputed** that "language data is converted into 'tokens' and that LLM models encode numerical values in a network of 'parameters' or weights." Books Plaintiffs' response consists of improper legal argument and recitation of additional facts about the "purpose of the training process" in the fair-use analysis that do not specifically controvert whether the LLM training process has a technical objective of "gain[ing] access to underlying facts, ideas, linguistic patterns, and statistical relationships." Books Plaintiffs' assertions do not specifically controvert that the steps described in Paragraph 19 are employed during the LLM training process for the objective of "gain[ing] access to underlying facts, ideas, linguistic patterns, and statistical relationships." For example, Plaintiffs do not dispute that training involves "tuning the LLM's parameters to encode the specific structure of different types of text." Plaintiffs' own expert agreed that through training, "the model will learn the most common patterns in the grammar, syntax and semantics of a given language," Hurst Decl. Ex. 370 (B. Zhao Tr.) at 29:15-24, and that a trained model "will exhibit some amount of

	generalization," ECF 1787 (Zhao Decl.) Ex. B ¶ 26; Hurst Decl. Ex. 370 (B. Zhao Tr.) at 338:3-13. Plaintiffs agree that the technical objective of training is predicting the next token in a sequence and that training LLMs on "a diverse array of textual genres and subject matter, foster[s] a more comprehensive understanding of language across various domains" and that "[i]t is widely accepted that to achieve high quality answers to a wide-ranging diverse set of user queries, LLMs must be trained with a diverse set of training data. ECF 1787 (Zhao Decl.) Ex. A ¶ 71. Defendants moved to exclude Zhao's "lossy compression" memorization opinion contained in for Zhao Decl. Ex. B ¶¶ 14, 25-27; Zhao Decl. Ex. B ¶¶ 26-27; Zhao Decl. Ex. A ¶ 12(a)(iv) pursuant to Federal Rule of Evidence 702, and so to the extent the Court excludes the challenged opinion, Plaintiffs may not rely on it to dispute this fact. ECF 2050; ECF 2055 at 3. Book Plaintiffs' characterization that "the purpose of training" uses "precise expression … and the expressive styles of different writers" is improper legal argument and expert characterization rather than a dispute of fact. Finally, Books Plaintiffs' assertion that "LLM training results in verbatim memorization" does not controvert the fact as stated for the reasons set forth in Microsoft's Reply to Paragraphs 13 & 14.

Paragraph 20

20. The patterns extracted through training—including grammar, syntax, semantic relationships, facts, ideas, and concepts—are properties of the aggregate training distribution rather than of any individual work. Lafferty Decl. ¶¶ 16, 33-34, 45, 66, 80-81, 99, 145-146. 283.	
Class Plaintiffs' Response: Disputed. As explained by Class Plaintiffs' experts, LLMs memorize individual works from their training data. Shan Decl. Ex. B ¶ 25 (explaining "the model has memorized the content of the work itself rather than copying or repairing its input"); Zhao Decl. Ex. B ¶ 72. Memorization of training data in an LLM means that during pre-training the LLM strongly associate the tokens making up a block of training text with one another as well as with references to that text. Zhao Decl. Ex. B ¶¶ 16, 26–27. Defendants knew that their LLMs memorized specific text from the training corpus. Nelson Decl. Ex. 964 at 3 ("The model will readily produce excerpts from works under copyright."); Nelson Decl. Ex. 965 at -776 (OpenAI employees stating that GPT-4 "will have memorized a ton of data and therefore will be insanely good at regurgitation"); Nelson Decl. Ex. 274 at -365 (OpenAI's Vice President of Research, Jerry Tworek: "We train our networks to memorize the training data - that's their objective."). **Further Disputed** because style is a pattern that LLMs can learn through training. Nelson Decl. Ex. 727 ¶ 71 (Microsoft's expert admitting that "the ability to generate news-like style or content is what one would expect from a model that has learned recurring statistical patterns of a genre"); Nelson Decl. Ex. 966 at 108:25-109:10 (former OpenAI employee confirming that, when a GPT model is presented with the first sentence or two of a work, the	**Microsoft's Reply: Undisputed** that LLMs learn patterns through training, and that those patterns include grammar, syntax, semantic relationships, facts, ideas, and concepts. Plaintiffs fail to genuinely dispute that "the patterns extracted through training … are properties of the aggregate training distribution rather than of any individual work." Plaintiffs' cited evidence does not controvert the fact as stated; it consists instead of additional facts concerning whether LLM training involves memorization of individual works and the extent to which an LLM is capable of regurgitation. As detailed in Microsoft's Reply to Paragraph 13, "memorization" is distinct from "regurgitation," and the "memorization" of individual works during LLM training is not inconsistent with patterns being derived from "the aggregate training distribution rather than of any individual work." Plaintiffs' evidence concerning "regurgitation" and what Defendant "knew" about regurgitation is flawed for the reasons stated in Microsoft's Reply to Paragraph 14, and does not show that patterns of grammar, syntax, semantic relationships, facts, ideas, or concepts are properties of any individual work. Further, Plaintiffs' own expert agreed that through training on a dataset, "the model will learn the most common patterns in the grammar, syntax and semantics of a given language." Hurst Decl. Ex. 370 (B. Zhao Tr.) at 29:15-24.

model could "generate something in a style that is similar to what it perceives to be the intended style or the kind of ongoing style of text"); Nelson Decl. Ex. 964 at 3-4 (June 2023 OpenAI presentation noting that "the model will readily produce excerpts from works under copyright" and that "it sometimes emulates the style").	Finally, Books Plaintiffs' assertion that "style is a pattern that LLMs can learn through training" does not controvert the stated fact and is improper legal argument that should be disregarded.

3. Impact Upon the Market (Factor 4)

Paragraph 21

21. Books Plaintiffs are authors whose traditional markets are publication by commercial publishers of books to be read in various formats such as ink on paper, e-books, and audiobooks, as well as the licensing of those commercially published books for translation into other languages and production in other entertainment mediums. Tucker Decl. ¶¶ 12(a), 25, 48, 64.	
Class Plaintiffs' Response: **Disputed** to the extent Microsoft limits the scope of licensing to licenses for "translation into other languages and production in other entertainment mediums." Admissible evidence contradicts Microsoft's statement. As set forth in greater detail in response to Microsoft's Statement of Fact No. 18, Class Plaintiffs who were asked testified that they would consider licensing their works for generative AI training, and an Authors Guild survey found that 95 percent of writers think they should be compensated for use of their works in developing generative AI. Lasinski Decl. Ex. B ¶ 238.	**Microsoft's Reply: Undisputed** that "Books Plaintiffs are authors whose traditional markets are publication by commercial publishers of books to be read in various formats such as ink on paper, e-books, and audiobooks" or that another "traditional market" for the Asserted Works is "the licensing of those commercially published books for translation into other languages and production in other entertainment mediums." Books Plaintiffs' assertion that "licensing" should also include "licensing their works for generative AI training" is improper legal argument and does not controvert the stated fact, which concerns "*traditional* markets" for their works. As detailed in Microsoft's Reply to Paragraph 18, testimony from a handful of authors that they would *consider* licensing their works for Generative AI training does not bear on whether that constitutes a "traditional market[]" that was excluded from the stated fact. To the extent that Plaintiffs contend that a non-traditional market for licensing their works for generative AI training exists, whether such a market "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.

Paragraph 22

22. Books Plaintiffs could not identify any lost sales or licensing opportunities for their Asserted Works as a result of the use of their Asserted Works to develop any of the LLMs at issue. Tucker Decl. ¶¶ 11(a), 20-46; *see also* Hurst Decl. Ex. 35 (Franzen Dep. Tr.) 46:5-8, 90:13-24, 91:1-22; Ex. 87 (Picoult Dep. Tr.) 93:5-16, 99:3-13; Ex. 39 (Golden Dep. Tr. (dated Mar. 17, 2025)) 210:16-20, 300:7-14, 301:20-302:2; Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 55:23-57:9; Ex. 128 (Wolicki Dep. Tr.) 148:17-151:20, 188:11-190:1; Ex. 108 (Shapiro Dep. Tr.) 91:15-92:4, 136:22-24, 223:6-17, 236:19-239:3; Ex. 9 (Baldacci Dep. Tr.) 224:20-225:10, 233:3-13, 238:1-15; Ex. 43 (Greer Dep. Tr.) 107:16-108:19, 110:2-19, 113:15-18; Ex. 73 (Martin Dep. Tr.) 216:22-218:6, 219:12-21; Ex. 55 (Hwang Dep. Tr.) 149:8-150:24, 151:16-152:25, 153:21-154:12; Ex. 106 (Schiff Dep. Tr.) 46:23-47:19; Ex. 44 (Grisham Dep. Tr.) 176:22-178:11.

Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. First, several of the Class Plaintiffs testified that they lost sales as a result of Defendants' unauthorized acquisition and use of Plaintiffs' works in training. Jodi Picoult testified that she lost "at least 27 sales" due to Microsoft's infringement. Nelson Decl. Ex. 927 at 93:5–10. Mary Rasenberger, CEO of the Authors Guild, testified that Defendants' piracy of the Authors Guild's works "were sales that should have happened" and "there is a loss" from Defendants' piracy. Nelson Decl. Ex. 967 at 17:10–16. Second, several of the Class Plaintiffs testified that they lost the opportunity to license their work to Defendants as a result of Defendants' unauthorized acquisition and use of Plaintiffs' works in training. David Baldacci testified that since Defendants "didn't come to [him] to license [his] work and they went to a pirate website, [he] lost the ability to license [his] work" and "lost those sales." Nelson Decl. Ex. 5 at 221:16–25. John Grisham testified that when Defendants "took [Plaintiffs'] protected property without	**Microsoft's Reply: Undisputed** that "Book Plaintiffs could not identify any lost sales or licensing opportunities for their Asserted Works *as a result of* the use of their Asserted Works to develop any of the LLMs at issue." Paragraph 22 cites to testimony from each author specifically addressing the stated fact. For example, Jonathan Franzen [redacted] Hurst Decl. Ex. 35 (J. Franzen Tr.) at 90:13-24, 91:1-22. John Grisham [redacted] Hurst Decl. Ex. 44 (J. Grisham Tr.) at 176:22-178:11. Every other Plaintiff testified likewise in the evidence cited in response to Paragraph 22. Additionally, Paragraph 22 cites an empirical analysis from Dr. Tucker demonstrating no

approval, without compensation, without licensing and used it for their LLMs, [he] lost money, as did every writer." Nelson Decl. Ex. 925 at 180:17–22. Sylvia Day confirmed that she lost a license to Microsoft Copilot because Microsoft used her works without placing a license "there [was] no license in front of [her]." Nelson Decl. Ex. 919 at 174:15–18. Mary Rasenberger, the Authors Guild corporate representative, testified that the Authors Guild "lost licensing income" for its works that defendants used as AI training data. Nelson Decl. Ex. 967 at 19:15–24.	decline in sales trends as a result of Defendants' products. ECF 1726 (Tucker Decl.) ¶¶ 11(a), 20-46. The evidence cited by Books Plaintiffs is insufficient to create a genuine dispute of fact. *See* Fed. R. Evid. 602, 701; Fed. R. Civ. P. 56(c)(4); *Argus Inc. v. Eastman Kodak Co.*, 801 F.2d 38, 45 (2d Cir. 1986) (self-serving testimony "consist[ing] largely of conclusory statements …., unsupported by documentary or other concrete evidence ….is simply not enough to create a genuine issue of fact in light of the evidence to the contrary"). For example, as detailed in Microsoft's Reply to Paragraph 25, Ms. Picoult testified that she *believed* she lost a single sale of each of her works used for training. However, when asked if she was aware of any documentation to support her assertion, she said there was "[redacted]" Hurst Decl. Ex. 87 (J. Picoult Tr.) at 93:11-16; *see also id.* at 94:11-14. The remaining testimony cited by Books Plaintiffs similarly consists of conclusory statements about potential lost sales or licensing opportunities that, when pressed, could not be supported by any concrete evidence to support their claims (as highlighted in the evidence cited by Microsoft in SUMF 22). *See, e.g.*, Hurst Decl. Ex. 98 (M. Rasenberger 30(b)(6) Tr.) at 22:17-22 [redacted]; Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 224:6-225:10, 233:3-13, 238:1-15 ([redacted]

	[redacted]); Hurst Decl. Ex. 44 (J. Grisham Tr.) at 176:22-178:11 [redacted]); Hurst Decl. Ex. 128 (S. Wolicki Tr.) at 148:17-151:20, 188:11-190:1 [redacted]. Such insufficient evidence does not create a genuine dispute of fact.

Paragraph 23

23. For 12 Asserted Works published in or after January 2017 (the most relevant time period), analysis of royalty statements and sales data shows the trend in unit sales and royalty earnings did not decline following the introduction of Copilot and ChatGPT; the observed trends reflect ordinary product-lifecycle dynamics. Tucker Decl. ¶¶ 11(a), 24-37.	
Class Plaintiffs' Response: **Disputed**. Admissible evidence contradicts Microsoft's statement. First, the royalty statements for several of the 12 Asserted Works Dr. Tucker identified in paragraph 11(a) of her declaration show a decline in royalties after the introduction of ChatGPT in November 2022 and the introduction of Copilot in 2023. Royalties for Andrew Sean Greer's novel *Less*, for example, declined from $99,780.85 in 2022 to $45,065.83 in the first half of 2023. Nelson Decl. Ex. 968 at -334; Nelson Decl. Ex. 969 at -407. The *Less* royalties then decreased from $45,065.83 in the first half of 2023 to $32,213.40 in the second half of 2023. *Id*. at -407; Nelson Decl. Ex. 970 at -457.	**Microsoft's Reply:** Paragraph 23 is not genuinely disputed. The fact states that analysis by Microsoft's expert for "12 Asserted Works published in or after January 2017" showed that "the *trend* in unit sales and royalty earnings did not decline following the introduction of Copilot and ChatGPT" and that "the observed trends reflect ordinary product life-cycle dynamics." The cited evidence does not controvert the fact as stated and contains improper legal argument concerning the weight to be given to the analysis. First, Plaintiffs point to a single instance where the *royalties* of one of the Asserted Works, *Less* by Andrew Sean Greer, declined from 2022 to 2023. However, Dr. Tucker

Second, Mr. Lasinski's analysis of Defendants' own regressions establishes statistically significant post-launch declines in unit sales of the Asserted Works. Reviewing the sales study performed by OpenAI's expert Dr. Gregory Leonard, Mr. Lasinski found that "[f]or July 2020, the Leonard Report reported a sample of 155 Asserted Copyrights, and found 71 of the 155 Copyright Works in sample had statistically significant negative trends in the post-period, which indicate a drop in unit sales," and that "[f]or November 2022, the Leonard Report documented a sample of 151 Copyrighted Works, and found 42 of the 151 Copyrighted Works in sample had statistically significant negative trends in the post-period, which indicate a drop in unit sales." Lasinski Decl. Ex. B ¶ 434. Mr. Lasinski likewise found that "0 of 67 Asserted Copyrights had an increase in trend that was statistically significant at the 95 percent confidence level," while "4 of 39 Asserted Copyrights had a decrease in trend that was statistically significant at the 95 percent confidence level." *Id.* ¶ at n.1457. As Mr. Lasinski concluded, that analysis "does not prove that none of the Copyrighted Works were harmed; some Copyrighted Works showed statistically significant decreases in sales trends in the post-period." *Id.* ¶ 435. Third, Mr. Lasinski's analysis refutes Microsoft's alternative explanation that "the observed trends reflect ordinary product-lifecycle dynamics." Mr. Lasinski found that when the Leonard Report "allowed for an extended post release timeframe, it found a larger proportion of statistically significant coefficients, suggesting that an impact might not be sufficiently captured within a 12-month window, and it does not establish the appropriate window of time," such that "data including a longer period	explained that *Less* experienced an increase in sales and royalties in 2022 due to the release of the subsequent installment in the series, *Less is Lost*. ECF 1726 (Tucker Decl.) ¶¶ 28-29. Academic research shows that such releases redirect consumer attention to prior titles, increasing their sales and royalties. *Id.* ¶ 29. Thus, a subsequent decrease in royalties for *Less* in 2023 following increased sales and royalties in 2022 reflects "ordinary product-lifecycle dynamics" and is "not consistent with a pattern of declining sales" that would be expected if the introduction of Copilot and ChatGPT negatively affected the trend in the sales and royalties of the Asserted Works. *Id.* Second, Plaintiffs attempt to dispute the stated fact, which concerns analysis done by *Microsoft's* expert, with their expert's critique of the separate analysis done by *OpenAI's* expert, Dr. Leonard. It is improper for Books Plaintiffs to rely on their expert's critique of OpenAI's expert as an undisclosed rebuttal opinion attempting to dispute a completely different analysis done by Microsoft's expert. Indeed, Dr. Leonard does not examine or even *discuss* Copilot. His analysis therefore does not dispute (and cannot since undisclosed) Dr. Tucker's findings regarding the lack of any adverse impact from the introduction of Copilot on the Asserted Works. Plaintiffs' undisclosed attenuated attack highlights the deficiencies in Plaintiffs' response and their lack of any evidence of actual harm in the traditional market. Third, even assuming that this critique of Dr. Leonard's analysis of ChatGPT is admissible against Microsoft's Copilot, it does not dispute this fact. Dr. Leonard performed *five* distinct "empirical analyses to evaluate Plaintiffs' claims of harm" and concluded that ChatGPT "had no

of time might be needed to properly identify the effect of Defendant's Accused Products on the Copyright Works' sales." Lasinski Decl. Ex. B ¶ 434. A decline that deepens as the post-release window lengthens is inconsistent with ordinary lifecycle decay, which would attenuate rather than intensify over time. Mr. Lasinski further identified an intervening market event that a lifecycle explanation does not account for: the "tripling of the number of ebook releases corresponding to the introduction of ChatGPT." *Id.* ¶¶ 375, 442; *see also* Lasinski Decl. Ex. C ¶ 86 (Amazon ASIN financial analysis). Fourth, the royalty and sales record on which Dr. Tucker relies is incomplete. Her analysis "entails summary statistics that do not address what the market would have looked like but for the alleged copyright infringement," and it omits "actual royalty data produced by the Plaintiffs" and "KDP internal documents produced by Amazon documenting ebook sales," as set forth in greater detail in response to Microsoft's Statement of Fact No. 24. Lasinski Decl. Ex. B ¶¶ 415, 442. As Mr. Lasinski explains, "[t]he failure to utilize ebook data from Amazon, the largest online bookstore in the world accounting for approximately 67 percent of ebooks sold worldwide, is particularly troubling given the Leonard Report's admission that 'AI-generated books are likely to be digitally distributed.'" *Id.* ¶ 415. Dr. Leonard similarly failed to evaluate any sales data for 50 Asserted Works and e-book sales for all of the Asserted Works. *Id.* An analysis that omits the royalty statements Plaintiffs produced and the channel through which AI-generated substitutes are principally distributed cannot establish the absence of a decline in unit sales or royalty earnings.	adverse impact on the sales of the Asserted Works" and "had no effect on the prices of trade books like the Asserted Works." ECF 1729 (Leonard Decl. ¶ 10). Plaintiffs ignore Dr. Leonard's full opinion and his conclusions, which takes into account all of his analyses, and instead cherry-pick not only results within these studies but also cherry-pick among the studies, only discussing two of the five. In discussing those two studies, Plaintiffs improperly note only the declines among Dr. Leonard's book study results, ignoring entirely the companion findings of *increases* and *no change*. In total, Leonard found there was no adverse impact from the at-issue LLMs on the sales of the Asserted Works in their traditional markets. *Id.* ¶¶ 10, 12-34. And *Mr. Lasinski himself* acknowledges that Dr. Leonard's mixed results are not causally associated with Copilot: "this analysis does not examine whether increases or decreases were caused by the introduction of Defendants' Accused Products or other supply and demand factors," ECF 1825 (Lasinski Decl.) Ex. B (Lasinski Rebuttal Rpt.) ¶ 435; *see also id.* ¶ 435, n. 1453 ("When a regression model fails to account for all factors influencing the dependent variable (*i.e.*, unit sales of the Copyrighted Works), the estimated coefficients will be biased, and one cannot draw meaningful conclusions[.]"). Indeed, half of Plaintiffs' last paragraph in its response is a *criticism* of those same analyses by Dr. Leonard. With results showing up, down and sidewise, and a conceded lack of causation from the product at issue, the fact of no change in historical trends is not disputed. Further, the Leonard selective results that Plaintiffs cite cannot dispute Tucker's findings regarding the works chosen during the period she deemed relevant. Two of Leonard's analyses use an event date of *July 2020*, which has no applicability to the fact as stated or the claims

	asserted against Microsoft Copilot. July 2020 "corresponds to when GPT-3 was made available via an API" and is *2.5 years before* the release of Copilot in February 2023. ECF 1729 (Leonard Opening Decl.) Exhibits 3a, 4a; ECF 1726 (Tucker Decl.) Figure 2 Notes (explaining that Bing Chat, which was later renamed Copilot, "was launched on February 7, 2023."). July 2020 is further unreliable as an event date because of the COVID impact on book sales during that time. As multiple Named Plaintiffs acknowledged, [redacted] Nelson Decl. Ex. 15 (Connelly Tr.) at 30:1-17; Hurst Decl. Ex. 418 (Baldacci Tr.) at 200:9-18 ([redacted]); *id.* at 201:3-6 [redacted]). Indeed, Mr. Lasinski acknowledged the importance of controlling for the impact of COVID when analyzing book sales, stating that certain analyses by Dr. Leonard "cannot be used to draw conclusions about economic harm" because they "fail[ed] to address or control for basic market considerations such as the impact of COVID[.]" ECF 1825 (Lasinski Decl.) Ex. B (Lasinski Rebuttal Rpt.) ¶ 431. In light of the July 2020 event date, Lasinski's critique has particular force against Plaintiffs' argument here that a downward trend is established, and the numbers they cite therefore "cannot be used to draw conclusions about economic harm." *Id.* Again, the finding by Dr. Tucker of a consistent trend in the traditional market is not disputed.

	Further, Mr. Lasinski's conclusion that an "extended post release timeframe" shows that "a longer period of time might be needed to properly identify the effect of Defendant's Accused Products" on the Asserted Works is entirely speculative and unsupported. A longer-time period allows for the possibility of more events that can impact a books' sales, like the series release bump, and then drop, described above for *Less* by Andrew Sean Greer. ECF 1726 (Tucker Decl.) ¶ 28; *see also id.* ¶ 31 (observing this effect in three other Asserted Works). Even assuming more books saw a decline in this "extended post release timeframe" that increase in decline cannot be attributed to Defendants' products because Mr. Lasinski fails to rule out other confounding factors or events that may have caused this decline. *See* ECF 1825 (Lasinski Decl.) Ex. B (Lasinski Rebuttal Rpt.) ¶ 435, n. 1453 ("When a regression model fails to account for all factors influencing the dependent variable (*i.e.*, unit sales of the Copyrighted Works), the estimated coefficients will be biased, and one cannot draw meaningful conclusions."). Regardless, that a *longer* time period might show different results does not dispute the findings of the specific time period evaluated in Paragraph 23. Fourth, a supposed "tripling of the number of ebook releases" since the introduction of ChatGPT does not dispute the stated fact. At most, it shows an increase in other books and is addressed to Plaintiffs' improper market dilution theory, not whether the 12 Asserted Works analyzed by Dr. Tucker in Paragraph 23 had any change from historical trends in sales or royalties. Finally, Plaintiffs' remaining assertions are merely an improper methodological attack that Plaintiffs failed to

	raise in a Rule 702 motion and should be disregarded. These criticisms—that "the royalty and sales record on which Dr. Tucker relies is incomplete" and that Dr. Tucker did not review Plaintiffs' royalty data, KDP data regarding ebook sales, or ebook sales of the Asserted Works—are both improper legal argument going to the weight of her opinions and are false. Dr. Tucker analyzed all of these data sources. For the 12 Asserted Works discussed in Paragraph 23, Dr. Tucker analyzed royalty statements produced by Plaintiffs that contained [redacted] ECF 1726 (Tucker Decl.) ¶¶ 24-25. Regarding the KDP data, sales of the 12 Asserted Works discussed in Paragraph 23 are not included in the KDP data. That data represents Kindle Direct Publishing, which is Amazon's self-publishing business. These 12 Asserted Works are not self-published. ECF 1726 (Tucker Decl.) ¶ 65. To the extent Plaintiffs argue that Dr. Tucker did not evaluate the *impact* of non-Plaintiff KDP ebook sales on Plaintiffs' books, that is also incorrect. Dr. Tucker analyzed the impact of a rise in the supply of self-published books on Amazon in 2018 and found no change in the daily average net unit sales of the Asserted Works. ECF 1726 (Tucker Decl.) ¶¶ 77-81. From this analysis, along with her review of the KDP Ditto data, Dr. Tucker concluded that the mere existence of additional self-published books does not mean that those books compete with the Asserted Works or take sales away from the Asserted Works. *Id.* ¶¶ 83-88. Plaintiffs' use of the term "AI-generated substitutes" is improper legal argument and is not supported by the record. ECF 1726 (Tucker Decl.) ¶¶ 11-12, 16-88.

Paragraph 24

24. For 125 Asserted Works published prior to January 1, 2017, the trend in sales also did not decline following the introduction of Copilot and ChatGPT. Tucker Decl. ¶¶ 11(a), 38-40.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. Dr. Catherine Tucker acknowledges that sales of the 125 Asserted Works declined after the introduction of ChatGPT and Copilot. *See* Tucker Decl. Figure 8 (showing decline); *see id*. ¶ 40 (explaining that "mean annual net unit" sales declined "3.5 percent in the year after the launch of Copilot" and had an "average decline of 8.2 percent per year over the two-year period following Copilot's introduction."). Likewise, OpenAI's expert, Dr. Gregory Leonard, found that 13 of the 21 Asserted Works with statistically significant trends had a negative trend in the period after the public release of ChatGPT. Lasinski Decl. Ex. B ¶ 433. Further, Dr. Tucker's analysis suffers from methodological shortcomings. Her "analysis entails summary statistics that do not address what the market would have looked like but for the alleged copyright infringement"; it "fails to account for basic considerations, such as the tripling of the number of ebook releases corresponding to the introduction of ChatGPT"; and it relies on a "Circana dataset," which "fails to account for actual royalty data produced by the Plaintiffs, KDP internal documents produced by Amazon documenting ebook sales," among other shortcomings. Lasinski Decl. Ex. B ¶¶ 415, 442. As Mr. Lasinski explains, "[t]he failure to utilize ebook data from Amazon, the largest online bookstore in the world accounting for approximately 67 percent of ebooks sold worldwide, is particularly	**Microsoft's Reply:** Paragraph 24 is not genuinely disputed. The fact states that analysis by Microsoft's expert also showed that "the trend in sales" for 125 Asserted Works published prior to January 1, 2017, did not decline following the introduction of Copilot and ChatGPT. The cited evidence does not controvert the fact as stated and contains improper legal argument. Plaintiffs' response selectively quotes from Tucker's declaration, which states in full: "the trend in physical sales for the [125] Asserted Works did not decline following the introduction of Copilot and ChatGPT: mean annual net unit physical sales [redacted] in the year prior to the launch of ChatGPT, compared to a smaller decline of 3.5 percent in the year after the launch of Copilot, and an average decline of 8.2 percent per year over the two-year period following Copilot's introduction." ECF 1726 (Tucker Decl.) ¶ 40. In other words, the rate of decline in sales for the Asserted Works after the introduction of ChatGPT and Copilot was *less* than the rate of decline prior to the introduction of these products. Books have ordinary lifecycles, so sales declines over the period after release are to be expected. A smaller decline obviously cannot be associated with harm. Additionally, Plaintiffs attempt to dispute this fact, which concerns analysis done by *Microsoft's* expert Dr. Tucker, with their expert Lasinski's critique of analyses done by *OpenAI's* expert, Dr. Leonard. Again, this is inadmissible as

troubling given the Leonard Report's admission that 'AI-generated books are likely to be digitally distributed.'" *Id.* ¶ 415.	undisclosed rebuttal opinion of Dr. Tucker, which she never had an opportunity to address. Indeed, Dr. Leonard does not examine or even *discuss* Copilot. His analysis therefore does not dispute (and cannot since undisclosed) Dr. Tucker's findings regarding the lack of any adverse impact from the introduction of Copilot on the Asserted Works. Even assuming that this critique of Dr. Leonard's analysis of ChatGPT is admissible against Microsoft's Copilot, it does not dispute this fact. Again, Plaintiffs cherry-pick one result within one of *five* empirical analyses conducted by Dr. Leonard, ignoring his overall conclusions drawn from his analyses. Plaintiffs improperly note only the declines among Dr. Leonard's book study results, ignoring entirely the companion findings of *increases* and *no change*. In total, Leonard found there was no adverse impact from the at-issue LLMs on the sales of the Asserted Works in their traditional markets. ECF 1729 (Leonard Decl.) ¶¶ 10, 12-34. And *Mr. Lasinski himself* acknowledges that Dr. Leonard's mixed results are not causally associated with Copilot: "this analysis does not examine whether increases or decreases were caused by the introduction of Defendants' Accused Products or other supply and demand factors," ECF 1825 (Lasinski Decl.) Ex. B (Lasinski Rebuttal Rpt.) ¶ 435; *see also id.* ¶ 435, n. 1453 ("When a regression model fails to account for all factors influencing the dependent variable (*i.e.*, unit sales of the Copyrighted Works), the estimated coefficients will be biased, and one cannot draw meaningful conclusions."). With results showing up, down and sidewise, and a conceded lack of causation from the product at issue, the fact of no change in historical trends is not disputed.

	Finally, Plaintiffs improperly attempt to dispute the fact with supposed "methodological shortcomings" in Dr. Tucker's analysis. With the exception of Plaintiffs' "summary statistics" criticism (with which Microsoft disagrees, *see* ECF 2357), these "methodological shortcomings" are merely an improper attack that Plaintiffs failed to raise in a Rule 702 motion and should be disregarded. These are improper legal argument going to the weight of her opinions and are false for the reasons stated in Paragraph 23. Plaintiffs' only additional criticism—that Dr. Tucker only analyzed Circana data, which focused on physical-format sales–ignores that Dr. Tucker conducted multiple analyses, some of which included physical and digital (ebook) formats and some of which focused only on physical formats. That the particular analysis cited in Paragraph 24 only focused on physical formats does not dispute the fact as stated. Additionally, the only ebook data available was Plaintiffs' royalty statements which were incomplete. ECF 1726 (Tucker Decl.) ¶¶ 24-25, 33. The Circana data provided a consistent and granular measure of physical sales trends over time that allowed Dr. Tucker to detect changes in physical sales levels or trends around the introduction of Copilot and ChatGPT that may not be visible in lower-frequency data, such as the semi-annual royalty statements produced by Plaintiffs. *Id.* ¶ 33.

Paragraph 25

25. The only existing and specific economic harm Books Plaintiffs have identified is the loss of payment for the single copy of each work used in LLM training. Hurst Decl. Ex. 44 (Grisham Dep. Tr.) 89:8-21; Ex. 106 (Schiff Dep. Tr.) 44:18-25, 45:11-46:6; Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 55:23-56:9, 58:6-13, 64:4-7; Ex. 87 (Picoult Dep. Tr.) 99:14-20; Ex. 128 (Wolicki Dep. Tr.) 149:22-150:5, 174:8-18; Ex. 9 (Baldacci Dep. Tr.) 223:3-14.

Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. Plaintiffs have identified other forms of economic harm, including market substitution and displacement of book sales. Plaintiffs' expert Michael Lasinski explained that authors are harmed by "the entry of a large quantity of new, competing content directly into the marketplace where Plaintiffs' works are sold," that the harm is amplified as "the introduction of AI-generated content meets or exceeds consumer acceptance levels," "thereby introducing a level of competition that did not exist before," and that "works can be created using AI quickly at lower cost and have lower prices than human-created works, on average," increasing "the likelihood of pricing pressure relative to copyrighted books." Lasinski Decl. Ex. B ¶ 411; *see also* Lasinski Decl. Ex. C ¶ 84 ("[Q]uality-driven harms and volume-driven harms – operate simultaneously in the marketplace for books, including copyrighted books, thereby providing adverse effects on authors of copyrighted books."); Seeley Decl. Ex. B ¶¶ 50-51 ("[G]enerative AI lowers barriers to entry in the market for copyrighted books and leads to a proliferation of AI content," and noting that Dr. Shan "generated more than 13,000 multi-chapter books in the span of two weeks using OpenAI's model"). Plaintiffs also identified harm in the form of lost licensing revenues that they otherwise would have received if Defendants had lawfully licensed their works for training and other purposes. Mr. Lasinski detailed "an emerging licensing market for acquiring access and use rights to data and content, including books, for training and/or output [which] has grown in scale and complexity." Lasinski Decl. Ex. B ¶ 53; *see also* Lasinski Decl. Ex. A ¶ 14 ("Defendants'	**Microsoft's Reply:** Paragraph 25 is not genuinely disputed. **Undisputed** that a harm identified by Books Plaintiffs "is the loss of payment for the single copy of each work used in LLM training." But Books Plaintiffs dispute that is the "only existing and specific economic harm" identified by Books Plaintiffs by identifying "other forms of economic harm." But Plaintiffs' cited evidence still fails to point to any "*existing* and *specific* economic harm" other than the loss of payment for a single copy used in training. Thus, the cited evidence does not specifically controvert the fact as stated and contains improper legal argument. Plaintiffs list general and speculative "harms" without providing any evidence that they have indeed suffered those harms. For example, Plaintiffs cite evidence that there has been an increase in AI-generated content generally but fail to identify any instance where a consumer purchased AI-generated content *instead of* one of the Asserted Works. Furthermore, Books Plaintiffs' 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." Gratz Decl. Ex. 116 (M. Lasinski Tr.) at 99:21-100:17. Mr. Lasinski testified the same was true with respect to Copilot. Hurst Decl. Ex. 414 (M. Lasinski Tr.) at 110:18-21. Regardless, 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* MSFT Books MSJ at 30-32. Plaintiffs' assertion that they have been harmed "in the form of lost licensing revenue" related to "licens[ing] their works for training and other purposes" is improper legal argument and is irrelevant to the fact as stated. Whether a market for

alleged wrongful conduct and infringement of the Copyrighted Works has negatively impacted the Plaintiffs in the form of lost revenues from . . . foregone licensing opportunities".). The viability of the licensing market for AI training and related purposes is "supported by Microsoft's own data initiatives and executed strategies, past practices and contract amendments, analyses related to content/data priorities and pricing, establishment of a licensing framework for negotiations, and real-world execution of data access and use agreements[.]" Lasinski Decl. Ex. B ¶¶ 104–13.

Plaintiffs identified further harm in the form of obstacles to authors' ability to monetize their works. Class Plaintiffs' expert Mark Seeley noted that researchers at the University of Cambridge found that "[o]ver a third (39%) of novelists reported that their income has already been negatively impacted by GenAI," which "was mostly attributed to competition from AI-generated books and loss of jobs which provide supplementary streams of income"; that "85% of published novelists said they think their future income will be negatively impacted by GenAI"; and that "over half (51%) of published novelists agreed that AI is likely to displace their work entirely." Seeley Decl. Ex. B ¶ 37 (citing Dr. Clementine Collett, *The Impact of Generative AI on the Novel*, at 8 (Nov. 2025), available at: https://www.mctd.ac.uk/wp-content/uploads/2025/11/MCTD-AIAndTheNovel-Report-Web.pdf). Likewise, Class Plaintiffs testified that AI-generated books and summaries have reduced demand for their works. Nelson Decl. Ex. 54 at 90:18–91:5.

Plaintiffs also identified reputational harm stemming from Defendants misconduct. Ms. Picoult testified, for example,

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. Regardless, Plaintiffs have also not identified any specific licensing opportunity they were not offered because of Defendants' conduct. *See* Hurst Decl. Ex. 87 (Picoult Tr.) at 99:14-100:2 ([redacted]); Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 224:2-225:10 (same); Hurst Decl. Ex. 43 (A. Greer Tr.) at 113:4-9 (same). Additionally, Plaintiffs have not shown that they sought to license their works for training or were approached by companies with offers to license their works for training. *See* ECF 1693 (Microsoft 56.1 Statement) 27(a) and 27(b). Thus, any claim of lost licensing revenue is speculative and non-specific and does not dispute the fact as stated.

Plaintiffs also identify potential "harm in the form of obstacles to authors' ability to monetize their works" but do not specifically identify how they have been unable to monetize their Asserted Works. To the contrary, the record shows that there has been no decline in the trend in unit sales or royalty earnings of Plaintiffs' Asserted Works. *See* ECF 1726 (Tucker Decl.) ¶¶ 16-60. In fact, the testimony of Ms. Picoult illustrates this exact point: [redacted] Hurst Decl. Ex. 87 (J. Picoult Tr.) at 93:5-10. Additionally, Mr.

that AI-generated outputs that are based on her style or use her characters, or which are inaccurate derivatives of her original works, can damage her reputation as an author. Nelson Decl. Ex. 927 (Picoult Tr.) at 207:5–208:12.	Seeley's mere recitation of statements in third party articles is hearsay. Thus, Plaintiffs' response to Paragraph 25 that relies on Mr. Seeley does not dispute the stated fact because it is not based on admissible evidence. Additionally, Plaintiffs claim that they have suffered "reputational harm stemming from Defendants misconduct." Yet the example they give from only *one* of the Plaintiffs, Ms. Picoult, highlights the speculative and non-specific nature of this alleged harm. She testified that [redacted] Nelson Decl. Ex. 927 (J. Picoult Tr.) at 207:13-14, 208:12. However, Ms. Picoult [redacted] Hurst Decl. Ex. 381 (J. Picoult Tr.) at 208:13-209:10. Regardless, Microsoft disputes that this alleged "reputational harm" is relevant to any element of copyright infringement, fair use, or cognizable market harm.

Paragraph 26

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

Class Plaintiffs' Response: Class Plaintiffs object to the phrase "market dilution" as vague and ambiguous, and further object to this statement to the extent it purports to state a legal conclusion regarding the scope of the fourth fair use factor. Subject to and without waiving those objections, **Disputed.** Class Plaintiffs incorporate by reference their responses to Microsoft's Statements of Fact Nos. 25 and 26(a) through 26(d). In addition, admissible evidence contradicts Microsoft's assertion. As Mr. Lasinski explains, the harm to authors operates through two concurrent market mechanisms—volume-driven entry of competing content and quality-driven consumer acceptance of that content—and both are measurable on this record. Lasinski Decl. Ex. B ¶¶ 410–11. The 2026 Reimers and Waldfogel study [redacted] demonstrate "that there has been a significant increase in the supply of books in the market, driven specifically by the proliferation of AI-generated content," *id.* ¶ 410, [redacted]	**Microsoft's Reply:** Paragraph 26 is not genuinely disputed. Microsoft's use of the phrase "market dilution" is not "vague and ambiguous"—Plaintiffs' opening brief in support of their motion for partial summary judgment states: "The first market harm is 'the harm of market dilution.'" BP MSJ at 34 (citing *Kadrey v. Meta Platforms*, 788 F. Supp. 3d 1026, 1052 (N.D. Cal. 2025)). Microsoft also incorporates by reference its replies to Book Plaintiffs' responses to Microsoft's Statements of Fact Nos. 25 and 26(a) through 26(d). As to the remainder of Plaintiffs' statement here, Microsoft addresses each below and incorporates its response by reference.

a. Books Plaintiffs have identified no instance in which an AI-generated book displaced a sale of any of Books Plaintiffs' commercially published Asserted Works; Hurst Decl. Ex. 35 (Franzen Dep. Tr.) 90:13-24; Ex. 108 (Shapiro Dep. Tr.) 90:15-92:4; Ex. 43 (Greer Dep. Tr.) 107:16-108:19; Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 47:21-48:6; Ex. 55 (Hwang Dep. Tr.) 150:10–24; Ex. 4 (Amer Dep. Tr.) 250:25–251:8; Ex. 9 (Baldacci Dep. Tr.) 231:5–12; Ex. 44 (Grisham Dep. Tr.) 187:8–13; Ex. 87 (Picoult Dep. Tr.) 94:16–23; Ex. 14 (Branch Dep. Tr.) 73:11–74:1; Ex. 73 (Martin Dep. Tr.) 219:13–21.

Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. Class Plaintiff David Baldacci testified that on the same day he released his book *Nash Falls*, AI-generated summaries of that book were sold "for a lower price" than his book. Nelson Decl. Ex. 5 at 227:3–228:6. Likewise, James Shapiro testified	**Microsoft's Reply:** Paragraph 26(a) is not genuinely disputed because the cited evidence does not controvert the fact as stated. Paragraph 26(a) cites to testimony from each author specifically addressing the stated fact, which concerns Books Plaintiffs' "market dilution" theory, and whether Books Plaintiffs have identified any instance in

that the "availability of these [AI-generated] summaries makes it easy for students who do not want to buy a recommended text, to shortcut that and not purchase my material, my book, my copyrighted material, but to allow you, through theft of my work, to profit from this. I think that does me significant harm." Nelson Decl. Ex. 54 at 90:18–91:5. Further, Mr. Lasinski opines that market displacement occurs by way of at least two concurrent market mechanisms. Lasinski Decl. Ex. B ¶ 411.	which a "AI-generated book" displaced a sale of an Asserted Work. For example, when asked whether "there are people out there who would have purchased digital or print copies of [the Asserted Works] but did not do so because of generative AI," David Henry Hwang testified that he has "no evidence to that effect." Hurst Decl. Ex. 55 (D. Hwang Tr.) at 151:3-9. David Baldacci [redacted] Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 231:11-12. George RR Martin testified that [redacted] Hurst Decl. Ex. 73 (G. Martin Tr.) at 219:13-21. Every other Plaintiff testified likewise in the evidence cited in response to Paragraph 26(a). Plaintiffs point to testimony from two Plaintiffs concerning "AI-generated *summaries*" of their own works, neither of which controverts the stated fact, which concerns "AI-generated *book[s].*" As reflected in Microsoft's cited evidence in support of Paragraph 23, however, those same two Plaintiffs testified that they could not identify any instance of an AI-generated book displacing sales of their own Asserted Works. *See* Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 231:11-12 (Baldacci testified that he had no [redacted]); Hurst Decl. Ex. 108 (Shapiro Tr.) at 90:15-92:4 (Shapiro confirmed that he "can't name anyone that didn't buy [his] book because of ChatGPT"). The testimony cited in Plaintiffs' response does not contradict these Plaintiffs' own testimony. David Baldacci's testimony that alleged AI-generated summaries of his book *Nash Falls* were sold "for a lower price" than his book does not establish that (1) any of these supposed

	summaries of *Nash Falls* were in fact AI-generated (as he [redacted], *see* Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 230:2-231:2, or (2) that any consumers purchased these supposed "AI-generated summaries" of *Nash Falls instead of* purchasing a full copy of *Nash Falls*. *See* Hurst Decl. Ex. 9 (D. Baldacci Tr.) at 231:5-9 [redacted]. The availability of AI-generated summaries (or books) on Amazon alone does not establish that they are actually being purchased or are competing with Plaintiffs' Asserted Works. ECF 1726 (Tucker Decl.) ¶¶ 61-88. The testimony of James Shapiro that Plaintiffs cite also does not controvert the stated fact for the same reasons. Similarly, Lasinski's opinion that "market displacement occurs by way of at least two concurrent market mechanisms" is vague, ambiguous, and does not establish that any consumer purchased an AI-generated book *instead* of one of the Asserted Works. Additionally, Defendants moved to exclude these "market displacement" opinions pursuant to Federal Rule of Evidence 702, and so to the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to dispute this fact. ECF 2050; ECF 2055 at 6-8. Because Plaintiffs offer no evidence that an AI-generated book actually displaced a sale of any of Books Plaintiffs' commercially published Asserted Works, this fact is not genuinely disputed.
b. Few if any people are using Copilot for the purpose of generating full books; Amir Decl. ¶¶ 9(b), 47-51; Bail Decl. ¶ 113; Tucker Decl. ¶ 12(d), 83-88; *see also* Hurst Decl. Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 62:17-63:2; Ex. 60 (Kazi Dep. Tr.) 143:19-144:1; Ex. 2 (Albu Dep. Tr.) 206:23-25; Ex. 43 (Greer	

Dep. Tr.) 145:12-15; Ex. 97 (Rasenberger Dep. Tr. (dated Feb. 12, 2026)) 282:18–23; Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 62:17-63:10; Ex. 4 (Amer Dep. Tr.) 252:17-25.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. [redacted] Microsoft's own survey data contains book-oriented uses. *See, e.g.*, Amir Declaration App'x H.1 (survey respondents indicating that they used Copilot to "[w]rite a story" and "create stories"). And Dr. Shan explained how "[g]enerating a full-length book from the production models is both fast and inexpensive." Shan Decl. Ex. D ¶ 58. Further **Disputed** because the analyses Microsoft cites were not designed to measure book-length generation and cannot support the conclusion Microsoft draws from them. Dr. Bail classified Copilot conversations against a fixed taxonomy of fifteen user-intent categories, and Microsoft identifies no category within that taxonomy corresponding to the generation of a full book. Bail Decl. ¶ 113. A classification scheme that defines no category for book-length output cannot establish that "[f]ew if any people are using Copilot for the purpose of generating full books." Nor can Dr. Amir's survey bear that weight: it measured only the task descriptions put to respondents, and even within those constraints respondents reported using Copilot to "[w]rite a story" and "create stories." Amir Decl. ¶¶ 47-51; Amir Decl. App'x H.1.	**Microsoft's Reply:** This fact is not genuinely disputed because the cited evidence does not controvert the fact as stated. Paragraph 26(b) cites to evidence from Microsoft's experts and Plaintiffs themselves specifically addressing the stated fact. For example, Dr. Amir found that only 1.8% of survey respondents indicated most recent Copilot uses that fell into the category Copilot for "writing / drafting of a book, poem, short story, or other literature" and 0.8% of survey respondents indicated using Copilot in the past twelve months for uses that fell into the same category. ECF 1716 (Amir Decl.) ¶¶ 45, 48. Ms. Albu (of Plaintiff The Authors Guild) was not aware of any authors who use Copilot in their writing. Hurst Decl. Ex. 2 (R. Albu Tr.) at 206:23-25. And when asked if he knew of a "specific example of a Microsoft Copilot-generated book published and offered for sale," Plaintiff Golden responded that he "couldn't speculate." Hurst Decl. Ex. 40 (C. Golden Tr.) at 62:21-63:2. The other testimony is similar. Plaintiffs' cited evidence does not controvert the stated fact. The data showing that [redacted] *See* Nelson Decl. Ex. 133 ( ) [redacted]. Additionally, this data does not establish that people used Copilot to "generat[e] full books" because [redacted] *See* Nelson Decl. Ex. 131 [redacted] at -288. Additionally, [redacted]

	[redacted] Nelson Decl. Ex. 761 ([redacted]) at ¶¶ 5, 6. Thus, the cited evidence does not support that any more than a "few" people are using Copilot to generate "full books." Plaintiffs' characterization of the findings of Drs. Bail and Amir are not only incorrect but also do not create a dispute of fact. Plaintiffs claim that none of the categories in Dr. Bail's classification schema correspond to "the generation of a full book." However, Dr. Bail's "Writing – Content Creation" was broad enough to capture the "generation of novels and other books." ECF 1719 (Bail Decl.) ¶ 113. And the evidence showed that people were not actually using Copilot for this purpose. *Id*. (noting that "in this category, conversations have on average 2.1 turns, with 68.9 percent of conversations consisting of a single turn, and 87.4 percent consisting of three or fewer turns. A user could not accomplish writing a novel that could compete with Plaintiffs' works using only a handful of turns."). Additionally, as noted above, Dr. Amir found that 0.8% of respondents indicated using Copilot for "writing / drafting of a book, poem, short story, or other literature," a category that was broad enough to cover "generating full books." ECF 1716 (Amir Decl.) ¶ 48. However, Dr. Amir's further analysis of the survey results made clear that generating a full book was not what the people in the .08% used Copilot for. Dr. Amir reviewed the 54 total responses that were coded under this category and found that a majority of respondents "described using Copilot to proofread or edit text they had

	already written, to experiment with creative writing prompts, or to draft short fictional passages for personal enjoyment or practice, rather than for uses that may have been derived from Class Plaintiffs' Asserted Works." *Id.* ¶ 49.
c. **AI-generated books are viewed with disfavor by consumers;** Tucker Decl. ¶¶ 12(b), 70-76, 87; *see also* Hurst Decl. Ex. 4 (Amer Dep. Tr.) 250:25-251:8; Ex. 87 (Picoult Dep. Tr.) 98:24-99:13; Ex. 35 (Franzen Dep. Tr.) 90:21-24; Ex. 55 (Hwang Dep. Tr.) 152:5-25; Ex. 44 (Grisham Dep. Tr.) 176:22-177:12; Ex. 43 (Greer Dep. Tr.) 108:16-19; Ex. 40 (Golden Dep. Tr. (dated Feb. 10, 2026)) 47:21-49:24; Ex. 60 (Kazi Dep. Tr.) 287:4-24; Ex. 97 (Rasenberger Dep. Tr. (dated Feb. 12, 2026)) 119:24-121:17.	
Class Plaintiffs' Response: **Disputed.** Class Plaintiffs object to the phrase "viewed with disfavor by consumers" as vague and ambiguous, and further object that the cited deposition testimony reflects individual authors' opinions about consumer preferences rather than evidence of consumer behavior. Subject to and without waiving those objections, **Disputed.** Admissible evidence contradicts Microsoft's statement. Consumers cannot reliably identify AI-generated writing: a 2023 study cited by Dr. Shan "showed that humans failed to identify AI writing approximately 50% of the time, and this study reflected text generated by LLMs weaker than, e.g., GPT-4o." Shan Decl. Ex. D ¶ 63. Consumer behavior likewise contradicts any claimed disfavor: [redacted] And as Mr. Lasinski explains, harm operates through a quality-driven mechanism precisely because "the introduction of AI-generated content meets or exceeds consumer acceptance levels," such that those works "can more directly compete with Plaintiffs' works, thereby introducing a level of competition	**Microsoft's Reply:** This fact is not genuinely disputed because the cited evidence does not specifically controvert the fact as stated. First, "viewed with disfavor by consumers" is not vague and ambiguous (as evident by Plaintiffs' testimony on this same subject). Second, authors' knowledge of their own preferences and their readers' preferences is relevant and admissible evidence of consumer behavior. The evidence cited by Plaintiffs does not contradict Microsoft's statement. First, the 2023 study cited by Dr. Shan only relates to whether consumers can *identify* AI-generated writing. It says nothing about whether they *prefer* AI-generated books. Moreover, that study applied to "self-presentation"-style writing in "professional, hospitality, and dating contexts" such as dating apps, as opposed to long-form fictional writing like the Asserted Works. Second, [redacted] *see* Nelson Decl. Ex. 133) has nothing to do with whether consumers view AI-generated books favorably; all this

that did not exist before." Lasinski Decl. Ex. B ¶ 411; *see also* Seeley Decl. Ex. B ¶¶ 50-51.	shows is that [redacted]; this statistic has nothing to do with how those KDP books containing some portion of AI-generated text performed or how many were sold. Finally, Mr. Lasinski's explanation "harm operates through a quality-driven mechanism" does not provide any evidence to dispute the fact as stated. Microsoft incorporates its response to Mr. Lasinski as stated in 26(a). Additionally, Microsoft notes that Plaintiffs mention a 2026 Reimers & Waldfogel study in response to Paragraph 26 above. As Plaintiffs pointed out in their opening 56.1 statement, that study 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." *See* BP SUMF ¶ 857.
d. **Books Plaintiffs concede they do not expect AI to replace them.** Tucker Decl. ¶¶ 45-46, 68-69; Hurst Decl. Ex. 25 (Connelly Dep. Tr.) 90:24-91:6; 160:3-13; Ex. 35 (Franzen Dep. Tr.) 90:21-24; Ex. 55 (Hwang Dep. Tr.) 152:5-25; Ex. 44 (Grisham Dep. Tr.) 176:22-177:12; Ex. 43 (Greer Dep. Tr.) 108:16-19; Ex. 87 (Picoult Dep. Tr.) 22:10-15; 98:24-99:13; Ex. 55 (Hwang Dep. Tr.) 150:10–24; Ex. 9 (Baldacci Dep. Tr.) 231:5–12.	
Class Plaintiffs' Response: **Undisputed** only that no Class Plaintiff testified that he or she expects artificial intelligence to replace authors entirely. **Disputed** to the extent Microsoft contends that the absence of such testimony bears on market harm, which does not require wholesale replacement. Admissible evidence contradicts any suggestion that Class Plaintiffs do not expect AI to	**Microsoft's Reply:** Paragraph 26(d) is not genuinely disputed. **Undisputed** that Class Plaintiffs testified they do not expect artificial intelligence to replace authors. Plaintiffs' attempt to manufacture a fact dispute by adding the word "entirely." But Books Plaintiffs do not offer any explanation of how that addition controverts the fact and does not cite any evidence to support this addition.

displace sales of and demand for their works. For example, Class Plaintiff John Grisham testified that "there is very real risk that we could be eliminated or seriously damaged by the flood of cheap books, legal thrillers." Nelson Decl. Ex. 24 at 175:1-17; *see also* Nelson Decl. Ex. 931 at 42:15–43:15 (testifying that "AI is producing, potentially long-term, millions of competitive works"); Nelson Decl. Ex. 927 at 53:19–54:14 (testifying that defendants' products allow generation of "more stuff that benefits them commercially but not me"); Nelson Decl. Ex. 54 at 67:11-20 (testifying that ChatGPT poses a threat to his future livelihood); Nelson Decl. Ex. 51 at 44:4-17 (testifying that she believed her future books were less likely to succeed).	Plaintiffs then "dispute" this statement by contending that it does not "bear[] on market harm." Plaintiffs' response is improper legal argument that does not dispute the stated fact. Additionally, the cited evidence does not controvert the fact as stated. Specifically, Plaintiffs cite the testimony of James Shapiro (Nelson Decl. Ex. 54 at 67:11-20) but omit that he explicitly testified [redacted] Nelson Decl. Ex. 54 at 67:17-18. Plaintiffs also cite to Jodi Picoult's testimony that Defendants used her works to "generate more stuff that benefits them commercially but not me" (Nelson Decl. Ex. 927 at 53:19–54:14) but this does not bear on the fact as stated: whether Plaintiffs expect AI to replace them. Instead, Ms. Picoult acknowledged that [redacted] Hurst Decl. Ex. 381 (J. Picoult Tr.) at 47:19-22. Additionally, when shown an AI output that was prompted to be written in the *style* of Jodi Picoult, she testified, [redacted] *Id.* at 234:25-235:1. The other testimony Plaintiffs cite may speak to Plaintiffs' belief that AI results in an increase of so-called AI-generated books, but Plaintiffs were not able to identify any instance where an AI-generated books displaced sales of their books. For example, Plaintiffs cite John Grisham's testimony about a speculated "flood of cheap books, legal thrillers" (Nelson Decl. Ex. 24 at 175:1-17) but he also testified [redacted]. Hurst Ex. 44 (J. Grisham Tr.) at 176:22-177:12. Plaintiffs cite Christopher Golden's testimony that "AI is producing, potentially long-

	term, millions of competitive works." (Nelson Decl. Ex. 931 at 42:15–43:15). But Mr. Golden was [redacted]. *Id.* at 44:7-22. And he was [redacted] *Id.* at 49:17-51:10. Lastly, Ms. Schiff testified that she is "not less likely to write a book" because of AI (Nelson Decl. Ex. 51 at 44:13-15) and she had no "quantifiable information" about book sales she lost because of AI. Nelson Decl. Ex. 51 at 47:19.

Paragraph 27

27. No agreements for licensing books as LLM training data existed at the time that the LLMs at issue were trained. Tucker Decl. ¶¶ 13(a), 90-91; *see also* Hurst Decl. Ex. 254 (2025.11.21 Hwang Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 255 (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 257 (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 258 (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 259 (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 293 (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; Ex. 260 (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 261 (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 253 (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 256 (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. [redacted] Nelson Decl. Ex. 971 ¶¶ 293-94, n.640; *see also* Nelson Decl. Ex. 885 at 43:4-7 (testifying that "there's an industry that provides training data to AI companies. It's actually been around for a very long time . . . ."); Lasinski Decl. Ex. B ¶¶ 138, 303–09 (referencing Meta licensing agreements granting AI	**Microsoft's Reply:** Paragraph 27 is not genuinely disputed. **Undisputed** that no agreements for licensing books as LLM training data existed prior to November 2022. The agreements cited by Plaintiffs, which were signed in or after November 2022, do not controvert the stated fact. As Plaintiffs acknowledge, OpenAI finished training GPT-3 in April 2020 and GPT-3.5 in December 2021. BP SUMF ¶¶ 521, 526. Additionally, the agreements cited by Plaintiffs concern textbooks and scientific books that are wholly different from Plaintiffs' Asserted Works—trade books.

training rights from early 2023 related to textbooks); ¶ 199 (referencing "[a] 2017 paper titled 'Artificial Intelligence's Fair Use Crisis' not[ing] that 'an established market for training data already exists.'").	[redacted] *See* Blythe Opp. Decl. Ex. 90 (OPCO_MDL_004600925) at -930 [redacted] *See* Nelson Decl. Ex. 971 (Leonard Opening Rep.) ¶¶ 293-96. [redacted] *Id.* Next, Plaintiffs cite Mr. Lasinski's hearsay discussion of "Meta licensing agreements granting AI training rights from early 2023 related to textbooks." ECF 1825 (Lasinski Decl.) Ex. B ¶¶ 138, 303–09. These appear to be references to two agreements listed in Mr. Lasinski's Rebuttal Attachment 6.0: [redacted]. ECF 1825 (Lasinski Decl.) Ex. B, Attachment 6.0, at (4) and (5). Even if Plaintiffs were citing the actual agreements (which they do not), these agreements are still inadmissible hearsay because Plaintiffs have not provided any authenticating or foundation evidence. Regardless, these agreements do not dispute the stated fact. [redacted]. *Id.* [redacted]

	[redacted] *Id.* [redacted] *Id.* [redacted] *Id.* Mr. Usdan's testimony (Nelson Decl. Ex. 885) does not controvert the fact as stated. His observation that "an industry that provides training data to AI companies" has "actually been around for a very long time" does not discuss *books* (just general "vendor data") and does not specify any particular timeframe in which such an industry existed. Additionally, when testifying about Microsoft's data access agreements related to LLM training (including agreements for books), Mr. Usdan testified that "there was no market. I mean, there is no market," and that "there's not a market for it." Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 325:2-11, 326:4-7. Finally, Lasinski's hearsay discussion of a 2017 draft paper that was not authenticated in this litigation constitutes hearsay and lacks foundation. In any event, Lasinski's report does not identify any "agreements for licensing books as LLM training data" identified in that 2017 paper, and his discussion of that paper therefore does not controvert the fact as stated. ECF 1825 (Lasinski Decl.) Ex. B ¶ 199.
a. **Books Plaintiffs have not offered any evidence that they were approached by any LLM developers to license their works for training;** Hurst Decl. Ex. 43 (Greer Dep. Tr.) 112:24-113:3; Ex. 128 (Wolicki Dep. Tr.) 157:16-18, 211:13-16; Ex. 106 (Schiff Dep. Tr.) 195:19-23; Ex. 14 (Branch Dep. Tr.) 108:21-109:6; Ex. 9 (Baldacci Dep. Tr.) 214:16-215:2; Ex. 87 (Picoult Dep. Tr.) 165:10-15; Ex. 44 (Grisham	

Dep. Tr.) 184:17-24; Ex. 254 (2025.11.21 Hwang Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 255 (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 257 (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 258 (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 259 (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 293 (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; Ex. 260 (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 261 (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 253 (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 256 (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.	
Class Plaintiffs' Response: Undisputed that Plaintiffs did not testify that they were directly approached by any LLM developers to license their works for training. **Disputed** to the extent Microsoft's statement is intended to suggest that Defendants did not desire or value Plaintiffs' books for purposes of developing LLMs. In February 2019, OpenAI indicated its willingness to pay for datasets containing books. Lasinski Decl. Ex. A ¶ 110 (citing Nelson Decl. Ex. 996). But OpenAI opted for piracy instead, because negotiating with authors and publishers would "involve protracted discussions / negotiations / licensing." Seeley Decl. Ex. A ¶ 47 (citing Nelson Decl. Ex. 998). Defendants' decision not to approach rightsholders was a deliberate choice to avoid the time and process of licensing, not a reflection of the works' value. OpenAI priced book acquisition as early as October and November 2018. Lasinski Decl. Ex. A ¶ 102. OpenAI nonetheless characterized monetary compensation options as "[o]perationally challenging + too slow," and Dario Amodei described acquiring books and papers through proper channels as "a huge legal/practice/business slog." *Id.* ¶¶ 107 n.523, 221 (citing Nelson Decl. Ex. 999 at -113).	**Microsoft's Reply: Undisputed** that Plaintiffs did not testify or otherwise offer evidence they were approached by any LLM developers to license their works for training. Plaintiffs' assertion regarding Defendants' alleged "desire or value" for their books is speculative, improper legal argument, and does not controvert the fact as stated. None of Plaintiffs' cited evidence establishes that any LLM developer actually approached Plaintiffs to license their works for training. Additionally, Plaintiffs' cited evidence is not admissible because it is hearsay and lacks foundation as to Microsoft because it relates to testimony, document, and alleged actions of OpenAI employees. Nelson Decl. Ex. 996 is a 2019 email chain between OpenAI employees. Nelson Decl. Ex. 998 appears to be a Slack thread between OpenAI employees. Nelson Decl. Ex. 999 appears to be a document produced by OpenAI in this litigation titled [redacted]

b. Books Plaintiffs have not offered any evidence that they sought to license their works for LLM training. Hurst Decl. Ex. 43 (Greer Dep. Tr.) 112:24-113:3; Ex. 106 (Schiff Dep. Tr.) 195:19-23; Ex. 98 (Rasenberger 30(b)(6) Dep. Tr.) 169:2-17; Ex. 14 (Branch Dep. Tr.) 108:25-109:6; Ex. 254 (2025.11.21 Hwang Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 255 (2025.11.21 Hieronymous Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 257 (2025.11.21 Golden Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 258 (2025.11.21 Gattopardo Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 259 (2025.11.21 Daring Greatly Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 293 (2025.11.21 Columbus Rose Supp. Resps. To Microsoft's 1st Set of Interrogs.) 7-8; Ex. 260 (2025.11.21 Belfrey Holdings Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 261 (2025.11.21 Wo & Shade LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8; Ex. 253 (2025.11.21 Sylvia Day LLC Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 6-8; Ex. 256 (2025.11.21 Greer Suppl. Resps. to Microsoft's 1st Set of Interrogs.) 7-8.

Class Plaintiffs' Response: Undisputed that Plaintiffs did not testify that they directly sought to license their works for AI training. **Disputed** to the extent Microsoft's statement is intended to suggest Plaintiffs' unwillingness to do so or a lack of a licensing marketplace. As set forth in greater detail in response to Microsoft's Statement of Fact No. 18, Plaintiffs testified that they would be willing to enter into a license for generative AI model training, and an Authors Guild survey found that 95 percent of writers think they should be compensated for use of their works in developing generative AI. Lasinski Decl. Ex. B ¶ 238 & n.861. Plaintiffs would have entered into such a license agreement had Defendants sought to negotiate and pay a fair price rather than take their works without permission. Numerous Class Plaintiffs have also participated in the settlement in *Bartz et al. v. Anthropic PBC*, No. 4:24-cv-05417-AMO (N.D. Cal.), further demonstrating their willingness to be compensated for the use of their works as AI training data. *See, e.g.*, Nelson Decl. Ex. 925 at 103:12–104:5; Nelson Decl. Ex. 931 at 53:20–54:15; Nelson Decl. Ex. 920 at 215:16–218:12.	**Microsoft's Reply: Undisputed** that Plaintiffs did not offer evidence they sought to license their works for LLM training. The evidence that Plaintiffs point to "that they would be willing to enter into a license for generative AI model training" does not specifically controvert the fact as stated. Specifically, the testimony set out in response to Microsoft's Statement of Fact No. 27 is from a handful of authors that they would hypothetically *consider* licensing their works for LLM training. None of the cited testimony shows that LLM training is a current type of license that any of the Books Plaintiffs have specifically sought out. Additionally, the Authors Guild survey asked writers generically whether they think they should be compensated for use of their works in developing generative AI. ECF 1825 (Lasinski Decl.) Ex. B ¶ 238. The survey does not specifically reflect the views of Plaintiffs, nor does it state that any author sought to license their works for LLM training. *Id.* Plaintiffs claim that "Plaintiffs would have entered into such a license agreement had Defendants sought to negotiate and pay a

	fair price rather than take their works without permission" but provide no support for this statement or evidence that any Plaintiffs did in fact seek to (or did) enter into such an agreement. That Plaintiffs opted-in to receive the *Bartz* settlement payment has no bearing on whether Plaintiffs sought to license their works for LLM training. A class action settlement agreement is not a content license transaction. *See* ECF 2222 (Tucker Opp. Decl.) ¶¶ 10-14. Further, the settlement is not "compensat[ion] for the use of [Class Plaintiffs'] works as AI training data" because the agreement release states that it covers claims "arising from Defendant's *alleged* past" conduct. Nelson Decl. Ex. 758 § 1.29 (emphasis added). The settlement agreement also 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." Nelson Decl. Ex. 758 § 1.29.

Paragraph 28

28. There are substantial impediments to the formation of a market for licensing a training corpus sufficient to train the type of general purpose LLMs at issue in this case.	
Class Plaintiffs' Response: **Disputed.** Class Plaintiffs incorporate by reference their responses to Microsoft's Statements of Fact Nos. 28(a) through 28(g) below. In addition, admissible evidence contradicts Microsoft's assertion. *See* Lasinski Decl. Ex. B ¶¶ 32, 159–167, 199 (explaining "the book marketplace, like marketplaces for other copyrighted works, possesses the institutional and contractual infrastructure that has enabled and continues to enable, adaptation to technological change"); Nelson	**Microsoft's Reply:** Microsoft incorporates by reference its replies to Book Plaintiffs' responses to Microsoft's Statements of Fact Nos. 28(a) through 28(g). Additionally, this fact is not genuinely disputed because the evidence cited by Plaintiffs does not controvert the fact as stated. Lasinski's discussion of the existing "book marketplace" does not controvert the stated fact because any licensing done in the existing "book marketplace" (i.e., print and adaption rights) is not comparable to licensing data for use

Decl. Ex. 972 at Tabs 13-14 (summarizing OpenAI, Microsoft, and Amazon's generative AI licensing agreements); Nelson Decl. Ex. 973 ¶ 130 (Microsoft's own expert admitting that "licensing collectives and marketplaces for data for AI uses" are "developing" in the United States); Lasinski Decl. Ex. A Attachment 6.0 (listing Microsoft's executed agreements); Lasinski Decl. Ex. A Attachment 6.1 (listing OpenAI's executed agreements). The scale of those transactions further refutes any claimed impediment: OpenAI's executed data and content access agreements have [redacted]	in LLM training, in part because no one content provider holds the rights to the large volume and variety of data needed for training purposes and value cannot be attributed to any particular piece of training data. *See* ECF 1726 (Tucker Decl.) ¶ 139. Additionally, the existence of one-off agreements (Nelson Decl. Ex. 972 at Tabs 13-14; ECF 1825 (Lasinski Decl.) Ex. A Attachments 6.0 and 6.1; ECF 1825 (Lasinski Decl.) Ex. B ¶¶ 132, 182) does not show either a lack of substantial impediments to the formation of a market or that such impediments are readily overcome. *Id.* ¶¶ 89-145. That OpenAI [redacted] and has entered agreements for access to data and content does not controvert that there are substantial impediments to the formation of a market for licensing a training corpus sufficient to train the type of general purpose LLMs at issue in this case. Additionally, the existence of [redacted] is not evidence that OpenAI has entered into any particular agreement, let alone one for trade books like the Asserted Works. ECF 1825 (Lasinski Decl.) Ex. B ¶ 182. Additionally, the fact that "OpenAI's executed data and content access agreements have [redacted] is not supported by the cited evidence. Neither paragraph of [redacted] includes this information. Regardless, that one developer has entered into any agreements does not controvert the fact. *See also* Hurst Decl. Ex. 382 (C. Tucker Tr.) at 250:25-251:8 ("[N]one of these agreements appear to solve these considerable economic frictions that I've identified in this hypothesized or hypothetical licensing market to exist, and indeed, when you look at the agreements, they are often covering multiple different other concerns, which again, makes it hard to interpret them as evidence in the way that the opposing experts are.").

	Lastly, Plaintiffs misstate Dr. Tucker's opinions regarding the status of licensing collectives and marketplaces for data for AI uses. Dr. Tucker does not conclude that functioning markets for licensing data for AI training or grounding are "developing," but only that certain experimental initiatives exist. ECF 2231 (Tucker News Opp. Decl.) ¶¶ 10(b), 18-24 (explaining that News Plaintiffs mischaracterize her testimony, and that recent initiatives are experimental and would still need to overcome considerable challenges to develop into a functioning market for licensing data for AI training or grounding at the scale required for general-purpose LLMs).
a. For example: A very large volume and a large variety of textual data is necessary for the training corpus to develop general purpose LLMs like those at issue; as such, no individual work has value by itself as training material nor is it currently technologically possible to isolate the contribution of any particular piece of data to an LLM's output generation; Lafferty Decl. ¶¶ 16, 34, 94-97, 99, 168, 283, 290, 322, 433; Usdan Decl. ¶¶ 10-11; Hurst Decl. Ex. 53 (Hoeschele Dep. Tr. (Vol. 1, dated Apr. 2, 2025)) 26:6-22, 35:18-36:11, 16:17-22, 17:1-23, 19:12-20:1, 23:24-24:25 25:15-23; Ex. 117 (Sutskever Dep. Tr.) 28:5-29:15, 67:9-15, 255:12-256:2.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. [redacted] Lasinski Decl. Ex. B ¶ 190. *Id.* Moreover, Defendants have in fact isolated the contribution of a category of training data to model performance: OpenAI tested	**Microsoft's Reply:** Paragraph 28(a) is not genuinely disputed. **Undisputed** that "a very large volume and a large variety of textual data is necessary for the training corpus to develop general purpose LLMs like those at issue." *See* MSFT Books SUMF ¶¶ 4-9, 12. Paragraph 28(a) then makes two additional points: (1) no *individual* work has standalone value as training material; and (2) it is not currently technologically possible to isolate the contribution of *any particular piece of data* to an LLM's output. Plaintiffs respond almost entirely with evidence about *aggregate* datasets, which is consistent with Microsoft's fact and does not contradict it.

whether public domain books could substitute for the LibGen books in its corpus and measured both a performance drop and a compute penalty, as set forth in greater detail in response to Microsoft's Statement of Fact No. 11. Nelson Decl. Ex. 289 at -367-73; Nelson Decl. Ex. 280 at -340; Zhao Decl. Ex. A ¶¶ 27-29. Further, Class Plaintiffs' expert, Dr. Ben Zhao, explained methods by which the value of LLM training data could be isolated. *See, e.g.*, Zhao Decl. Ex. A ¶ 64 (describing "a methodology known as machine unlearning," which has been used to identify "books as high-impact data with disproportionate influence on model capabilities relative to their share of training tokens"). OpenAI showed that it recognized such value by increasing "midtraining" the proportion of books used in its training datasets. *Id.* ¶ 36 (citing Nelson Decl. Ex. 352). Microsoft's own personnel have likewise stated that it is possible to isolate the contribution of a particular piece of data to an LLM's output generation. Nelson Decl. Ex. 974 at -942 (Microsoft's Jaron Lanier proposing in March 2023 that Microsoft "trace the most unique and influential contributors of prior examples or data when a big AI model provides a valuable output" in order to "calculate[]" which data contributors were "uniquely essential" so they "would get paid").	First, Plaintiffs' cited evidence does not controvert that "no individual work has value by itself as training material." Mr. Lasinski's observation that [redacted] ECF 1825 (Lasinski Decl.) Ex. B ¶ 190, is consistent with the stated fact and is merely a description of Defendants' approach. Plaintiffs' citation to Microsoft's data access agreements also does not controvert the stated fact. *See* ECF 1707 (Usdan Decl.) ¶ 11 (Mr. Usdan, Microsoft's 30(b)(6) witness on the purchase of data for LLM training who was involved with these agreements, stated that his "team went in circles trying to arrive at a way of valuing data for AI training, including consulting scientists like [Microsoft employee] Mr. Bubeck, but it is extremely hard to determine the value of a given piece of data is [sic] to a model, and then the value, if any, of any increase of performance of that model based on that data input." ECF 1707 (J. Usdan Decl.) ¶ 11. Mr. Usdan further stated that Microsoft's agreements [redacted]. *Id.* ¶¶ 25, 28. Second, Plaintiffs' cited evidence does not create a dispute on whether it is "technologically possible to isolate the contribution of any particular piece of data to an LLM's output generation." Microsoft incorporates by reference its Reply to Paragraph 11. Plaintiffs mischaracterize the evidence cited as Nelson Decl. Ex. 974, which appears to reflect personal thoughts by an individual Microsoft

	employee concerning the theoretical concept of "data dignity" and cannot be attributed to Microsoft. The quoted language discusses a *philosophical* desire to compensate contributors to AI and does not establish that it is *technologically* possible to isolate any such contribution. Nelson Decl. Ex. 974 at -942. Pursuant to Federal Rule of Evidence 106, Mr. Lanier testified that there is no way to do it technologically: "we don't really know" that "if you remove some stuff from training data, how much would it affect the output from time to time." Hurst Decl. Ex. 369 (J. Lanier 30(b)(1) Tr.) at 264:3-9. He also testified that "the way a large language model works is reflections of reflections of reflections, all statistical, and so there isn't a single place where you can say this is the thing and this is how it's used.... you can remove it and then it will still reconstruct itself from all the reflections out there and other data." *Id.* at 107:8-17. This supports the fact that there is no way to value any specific data for training or outputs. The other evidence cited by Plaintiffs on this point is not admissible because it is hearsay and lacks foundation as to Microsoft because it relates to testimony, document, and alleged actions of OpenAI employees. Nelson Decl. Exs. 289 and 280 appear to be Slack threads between OpenAI employees. Zhao Declaration Ex. A ¶¶ 27-29 also discusses OpenAI documents and OpenAI employee testimony. *See* ECF 1787 (Zhao Decl.) Ex. A ¶¶ 27-29. Dr. Zhao's explanation that "methods by which the value of LLM training data could be isolated" also do not controvert the fact as stated. His "machine unlearning" methodology describes a research technique that he did not do. It is not evidence that the contribution of any particular work to the Accused Models' outputs can currently be isolated. Additionally, Nelson Decl. Ex. 352 (discussed in Zhao Decl.

	Ex. A ¶ 36) is inadmissible hearsay and lacks foundation as to Microsoft because it relates to an OpenAI document.
b. The Asserted Works were a minuscule share of the training corpus used to train the GPT models; Tucker Decl. ¶¶ 13(b), 95, 99-100; Lafferty Decl. ¶¶ 94-97, 105-109.	
Class Plaintiffs' Response: **Disputed.** Class Plaintiffs object to the phrase "minuscule share" as vague and ambiguous. Subject to and without waiving that objection, **Disputed** because Microsoft offers no evidence quantifying the share of any training corpus attributable to the Asserted Works, and because share of tokens does not measure contribution: copyrighted books, including the Asserted Works, were an especially important part of the training corpus used to train the GPT models. As explained by Class Plaintiffs' expert Michael Lasinski, "[w]hen training GPT-3, the Books1 and Books2 datasets had a combined 'weight in training mix' greater than their token quantities alone would suggest," and "datasets containing Copyrighted Works (LibGen and Common Crawl) accounted for [redacted] of the weight of the training mix of GPT-3." Lasinski Decl. Ex. C ¶ 45; *see also* Nelson Decl. Ex. 869 at 59:16–61:25 (noting an OpenAI benchmark study describing LibGen data of "books and novels" as a [redacted]); Nelson Decl. Ex. 168 (benchmark study referenced in Parakhin's testimony). In a May 2020 internal writeup shared with Microsoft, OpenAI identified its training dataset as consisting of four sources—"Commoncrawl (filtered)," "Webtext," "LibGen," and "Wikipedia"—with "LibGen" alone encompassing [redacted] of the "training mix." *Id.* at -654. OpenAI further explained to Microsoft that it [redacted] "so that they	**Microsoft's Reply:** Paragraph 28(b) is not genuinely disputed. Books Plaintiffs' cited evidence does not controvert this fact. Specifically, one of Plaintiffs' own SUMFs recognizes that "GPT3 was trained on the equivalent of [redacted] books" (BP SUMF ¶ 523), whereas Plaintiffs have only asserted 212 works in this case. Thus, there can be no genuine dispute that the Asserted Works (to the extent they were even used to train the GPT models) were a "miniscule share of the training corpus." (212/[redacted]). The evidence cited by Plaintiffs in response only relates to GPT-3, and no other at-issue models. Regardless, even that evidence does not dispute the fact as stated. Mr. Lasinski's assertion that datasets allegedly containing LibGen and Common Crawl material accounted for [redacted] of the "weight" of the training mix does not dispute this fact, because it fails to offer any estimate of a "share" of the *Asserted Works* in the training corpus. Plaintiffs' reliance on Mr. Parakhin's testimony regarding a "benchmark study" likewise does not dispute this fact, as that merely describes "books and novels" as [redacted] Nelson Decl. Ex. 869; Nelson Decl. Ex. 168. That evidence does not relate at all to the share of the Asserted Works in the training corpus. Likewise, the discussion of the alleged share of "Libgen" in the "training mix" says nothing about the share of the *Asserted Works* in the training data. *See* Nelson Decl. Ex. 168; Nelson Decl. Ex. 697.

occurred [redacted] of the time during training," including "a dataset from Library Genesis (Libgen), which contains books and novels [redacted]," with the result that [redacted] Nelson Decl. Ex. 697 at -056.	
c. **For a traditional copyright license, rightsholders monetize the inherent value of specific creative expression contained within their individual works. By contrast, when licensing data for LLM training, value is not attributable to any discrete piece of interchangeable data, but rather emerges from statistical distributions observable only across large-scale datasets. There is no settled or workable means for pricing any given piece of training data;** Tucker Decl. ¶¶ 13(c), 111-114.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. Defendants' own experts opined that pricing training data is workable. Microsoft's expert Dr. Mohan Rao opines that a "Market Approach" is an appropriate way to value books used for training. Nelson Decl. Ex. 972 ¶ 170. OpenAI's expert Maureen Chakraborty opines that "OpenAI materials show that OpenAI [redacted] Nelson Decl. Ex. 975 ¶ 32. And pricing structures across independent transactions are converging on recognizable per-token, per-query, per-scrape, revenue-share, and lump-sum conventions. Nelson Decl. Ex. 204 at -089 [redacted]); *see also* Lasinski Decl. Ex. A Attachments 6.0, 6.1 (listing Microsoft's and OpenAI's executed agreements). [redacted]	**Microsoft's Reply:** Paragraph 28(c) is not genuinely disputed. **Undisputed** that "[f]or a traditional copyright license, rightsholders monetize the inherent value of specific creative expression contained within their individual works." Plaintiffs' cited evidence does not controvert (1) that LLM licensing value "emerges from statistical distributions only observable across large-scale datasets" or that (2) "[t]here is no settled or workable means for pricing any given piece of training data." None of the cited evidence shows a different assignment of value or a reliable method for pricing any one work's contribution to model training. At most, Plaintiffs' cited evidence shows that the parties' damages experts were able to construct a lost license fee solely for damages purposes and that the parties were able to reach agreement in certain specific circumstances for the use of content in LLM training. That does not controvert the fact stated in Paragraph 28(c). Dr. Rao and Dr. Chakraborty's reliance on *past* agreements where parties were able to reach agreement on price in individual circumstances for *a damages analysis in litigation* is not evidence that there is a "settled or workable

[redacted] Nelson Decl. Ex. 976 at -554–55.

Class Plaintiffs' experts likewise opine that pricing books as training data is workable, based on Defendants' own licensing practices. Lasinski Decl. Ex. B ¶ 190; *see also* Zhao Decl. Ex. A ¶ 26 (noting that OpenAI employee Qiming Yuan benchmarked the value of a "high-quality, diverse e-books dataset" at "10X of the cost of training a model").

means for pricing any given piece of training data." Indeed, Dr. Rao specifically noted Dr. Tucker's finding that there are "difficulties inherent to pricing or attributing value to any piece of content" for training LLMs. Nelson Decl. Ex. 972 at ¶ 169. Additionally, as Plaintiffs point out in their response, that there are multiple pricing structures across transactions, such as "per-token, per-query, per-scrape, revenue-share, and lump-sum conventions" *supports* that there is no "*settled* or workable means for pricing any given piece of training data."

The Microsoft presentation cited as Nelson Decl. Ex. 204 likewise is not evidence that there is a settled "pricing structure" convention. Mr. Usdan testified that presentation was "in the very early stage of trying to get our heads around what would be a fair price to pay for a book" and certain pricing ranges were not based on any value assigned to an individual book but were based on Mr. Usdan "trying to reverse engineer a justification for this huge, ridiculous price we were being quoted." Hurst Decl. Ex. 359 (J. Usdan 30(b)(6) Tr.) at 266:13-15. Mr. Usdan further explained that [redacted] *Id.* at 267:16-20. Regardless of what this early presentation said, [redacted]

[redacted]" ECF 1707 (Usdan Decl.) ¶ 24(b).

Likewise, Nelson Decl. Ex. 976 similarly reflects [redacted] does not controvert that there is no "settled and workable means" to value data for LLM training; it proves it. Plaintiffs' reliance on Dr. Zhao's hearsay recitation of a statement by an OpenAI employee fails for the same reasons—it merely reflects discussion and experimentation on how to hypothetically value data; it does not provide a workable means to value data. *See* ECF 1787 (Zhao Decl.) Ex. A ¶ 26.

Plaintiffs' reliance on agreements entered into by Microsoft and OpenAI fares no better. *See* ECF 1825 (Lasinski Decl.) Ex. B ¶ 190; ECF 1787 (Zhao Decl.) Ex. A ¶ 26. Those agreements do not provide support to dispute the fact as stated. First, [redacted], and so the value of any one agreement is not directly tied to the value of data to training [redacted] ECF 1726 (Tucker Decl.) ¶¶ 128-129. That two parties agreed to a price for a particular transaction does not necessarily reflect an agreed-upon value of data for LLM training, and could reflect those many other considerations. *Id.* Additionally, just because certain agreements have included

	a particular price does not mean that there is a "settled or workable means for pricing any given piece of training data" or that books actually provide value to LLM training. To this point, Jordan Usdan has explained that there is "no accepted or reliable mechanism for valuing how one data input affects AI training." ECF 1707 (Usdan Decl.) ¶ 11.
d. A developer seeking to license a training corpus would need to obtain rights from millions of dispersed authors and content creators; Tucker Decl. ¶¶ 13(c), 89, 101-106; Usdan Decl. ¶¶ 10-11, 22, 24.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. The purported difficulties are not especially hard to surmount. LLM developers did not have to deal with millions of authors at one time, and they did not attempt to do so. Instead, [redacted] *See, e.g.*, Nelson Decl. Ex. 971 ¶¶ 293–94 (referencing [redacted]; Lasinski Decl. Ex. A ¶ 110 (noting [redacted]); Lasinski Decl. Ex. B ¶¶ 168–174 (noting that initial contracts with publishers "had to be negotiated" but plan was to "utilize a standardized contract in the future"); Nelson Decl. Ex. 977 at -585 (Greg Brockman of OpenAI stating that [redacted]. Moreover, intermediaries also exist to facilitate licensing of content for generative AI. *See* Lasinski Decl. Ex. B ¶¶ 59–103 (describing the intermediaries, technological solutions, and collective licensing solutions that have entered the	**Microsoft's Reply:** Paragraph 28(d) is not genuinely disputed because the evidence cited by Plaintiffs does not controvert the fact as stated. The evidence Plaintiffs cite does not show that a developer can license a training corpus without obtaining rights from millions of dispersed authors and content creators. Instead, the cited evidence shows that it is possible to obtain rights from *some* authors, which does not dispute the fact as stated. First, Plaintiffs claim that the existence of publishers disputes this fact. But the ability to license through publishers does not necessarily relieve an LLM developer from having to negotiate with millions of authors and content creators to license a training corpus. Even if a developer negotiated an agreement with a publisher, the publisher may still need to get permission from individual authors to license their works for LLM training. *See, e.g.*, ECF 1707 (Usdan Decl.) ¶ 22 ("[redacted]"). Moreover, individual publishers remain sufficiently dispersed that any one agreement would still only encompass a miniscule fraction of the content

market to facilitate licensing in the generative AI data licensing market).	needed to train a general purpose LLM. *See* Paragraph 28(b). For example, [redacted] cited by Plaintiffs, which is not for trade books and does not contain any financial term by which OpenAI paid Springer Nature for content [redacted] *See* Blythe Opp. Decl. Ex. 90 (OPCO_MDL_004600925) at -932. For comparison, the Library of Congress holds about 25 million books, which is equivalent to approximately 2.5 trillion tokens. *See* ECF 1726 (Tucker Decl.) ¶¶ 95-96. [redacted] *Id.* Microsoft also objects to Plaintiffs' reliance on Nelson Decl. Ex. 977, which is an OpenAI Slack channel, because it is hearsay and lacks foundation as to Microsoft. Second, the mere existence of the intermediaries cited by Plaintiffs does not necessarily obviate the need to negotiate with millions of individual authors. ECF 1726 (Tucker Decl.) ¶¶ 131-136. These intermediaries are still in experimental stage and typically operate on opt-in designs, involve a limited subset of rightsholders, and primarily concern rights to display copyrighted content, not train an LLM. *Id.* ¶ 131. Plaintiffs have not cited any evidence that intermediaries could provide the rights for the amount of data needed to train an LLM.
e. Rightsholders for individual works can be difficult and costly to identify and locate, increasing the transaction costs of negotiating any proposed training data licenses; Tucker Decl. ¶¶ 13(c), 102-106; Usdan Decl. ¶ 24.	
Class Plaintiffs' Response: Disputed. Admissible evidence contradicts Microsoft's statement. As explained by Class Plaintiffs' expert Michael Lasinski, "the traditional	**Microsoft's Reply:** Paragraph 28(e) is not genuinely disputed because the evidence cited by Plaintiffs does not controvert the fact as stated.

book market has repeatedly expanded its licensing frameworks to accommodate innovations such as e-books and audiobooks." Lasinski Decl. Ex. B ¶¶ 12, 28–29, 159–167. Transaction costs have not in fact impeded licensing: Microsoft's and OpenAI's [redacted], *see* Lasinski Decl. Ex. A Attachments 6.0, 6.1; *see also* Lasinski Decl. Ex. B ¶¶ 180–183. For example, [redacted]. Nelson Decl. Ex. 975 App'x E ¶ 3(a) (referencing OpenAI partnership agreement with Springer Nature). Plaintiffs in *Bartz et al. v. Anthropic PBC*, No. 4:24-cv-05417-AMO (N.D. Cal.) were able to locate hundreds of thousands of rightsholders during the settlement administration process, resulting in a claims rate of nearly 93 percent as of April 16, 2026. Lasinski Decl. Ex. B ¶ 157 & n.580. Public registries, like Amlet, also exist "allowing users to recognize a work [and] verify the rights holders," and "when for-profit enterprises have attempted to create a new revenue stream for book copyright holders, all parties to the licensing negotiation are able to efficiently identify themselves and reach agreement." *Id.* ¶¶ 68, 167.	First, Lasinski's discussion of the "traditional book marketplace" does not dispute that there are transaction costs associated with negotiating proposed training data licenses. Licensing in the "traditional book marketplace" (i.e., print and adaption rights) is not comparable to licensing data for use in training, in part because no one content provider holds the rights to the large volume and variety of data needed for training purposes and value cannot be attributed to any particular piece of training data. *See* ECF 1726 (Tucker Decl.) ¶ 139. The existence of one-off executed agreements (ECF 1825 (Lasinski Decl.) Ex. A Attachments 6.0 and 6.1, *see also* ECF 1825 (Lasinski Decl.) Ex. B ¶¶ 180–183) does not dispute that transaction costs would have to be overcome at the scale large enough to acquire the necessary amount of data needed to train an LLM. Plaintiffs have not identified any evidence of this occurring. Plaintiffs' citation to [redacted] further illustrates this point, as that reflects a miniscule portion of the content needed to train a general-purpose LLM. *See* Microsoft's Reply to Paragraph 28(d). Additionally, agreements that do exist are further evidence that the transaction costs for such agreements are high. Jordan Usdan explained that [redacted]

	[redacted] ECF 1707 (Usdan Decl.) ¶ 24(b). Plaintiffs' reliance on the *Bartz* settlement also does not dispute the fact as stated. First, that authors opted in to receive the *Bartz* settlement payment has no bearing on the transaction costs for licensing content at scale to train a general-purpose LLM. A class action settlement agreement is not a content license transaction. *See* ECF 2222 (Tucker Opp. Decl.) ¶¶ 10-14. That settlement, which includes a work list of 482,460 works, still represents only a miniscule share of the content needed to train a general purpose LLM. *Id*. ¶ 10; *see also* Paragraph 28(b). Regardless, the *Bartz* settlement is indicative of the transaction costs and difficulties in identifying rightsholders. *See* ECF 2222 (Tucker Opp. Decl.) ¶ 11(b). Indeed, "on repeated occasions, authors and publishers have staked competing claims with the settlement administrator[.]" *Id*. (quoting New York Times article). Mary Rasenberger, CEO of the Authors Guild, explained in a *Publishers Weekly* article that "not all publishers keep great records of what books they've reverted rights to." *Id*. Additionally, the existence of public registries of works does not dispute that transaction costs of negotiating any proposed training data licenses exist. Registries like Amlet still require rightsholders to register their works and do not guarantee that the correct rightsholder can in fact be located.

f. Negotiations for the aggregation of the necessary volume and variety of data are susceptible to holdout problems; Tucker Decl. ¶¶ 13(c), 107-110.

Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement. Class Plaintiffs' expert Michael Lasinski explains that "the Tucker Report's discussion of holdout problems is speculative, as she has not identified any alleged instances of holdout related to the licensing market for copyrighted books for LLM training data. Indeed, the facts of this case demonstrate that licensors are willing participants in the market for licensing or accessing data for LLM training." Lasinski Decl. Ex. B ¶ 197; *see also* Lasinski Decl. Ex. A Attachments 6.0, 6.1 (listing Microsoft's and OpenAI's executed agreements).	**Microsoft's Reply:** Paragraph 28(f) is not genuinely disputed because the evidence cited by Plaintiffs does not controvert the fact as stated. The limited number of agreements that Microsoft and OpenAI have entered into support that holdout problems exist. Indeed, [redacted] LLMs. Blythe Opp. Decl. Ex. 77 (J. Usdan 30(b)(6) Tr.) at 325:12-24. In addition, the existence of one-off executed agreements (ECF 1825 (Lasinski Decl.) Ex. A, Attachments 6.0 and 6.1) does not dispute that holdout problems would have to be overcome at the scale large enough to acquire the necessary amount of data needed to train an LLM. ECF 1726 (Tucker Decl.) ¶¶ 13(c), 107-110.

g. When Microsoft licensed book data from [redacted] for LLM training at a price the publisher was willing to accept, the publisher [redacted], and Microsoft ultimately declined to continue that agreement or to pursue any others. Usdan Decl. ¶¶ 22-25.

Class Plaintiffs' Response: **Undisputed** that Microsoft licensed book data from [redacted] for LLM training, that [redacted] **Disputed** as to the balance of the statement, including that [redacted]	**Microsoft's Reply:** Paragraph 28(g) is not genuinely disputed. **Undisputed** that "Microsoft licensed book data from [redacted] for LLM training" and that "Microsoft ultimately declined to continue that agreement." The remainder of Books Plaintiffs' response does not create a genuine dispute because the cited evidence does not controvert the stated fact.

[redacted]	First, Plaintiffs purport to dispute that Microsoft declined to "pursue any other[]" agreements like [redacted] that Plaintiffs cite and thus does not dispute the fact as stated. Microsoft employee Jordan Usdan stated: [redacted] ECF 1707 (Usdan Decl.) ¶ 25. Additionally, Microsoft [redacted] *See* Defs.' Resp. to BP SUMF ¶ 811. Mr. Suleyman's testimony does not identify any deals [redacted] and therefore does not controvert the stated fact. Nelson Decl. Ex. 980 at 147:6-148:19. In fact, Mr. Suleyman was [redacted] . *Id.*
Admissible evidence contradicts Microsoft's statement. [redacted] Nelson Decl. Ex. 980 at 145:15-147:16; Nelson Decl. Ex. 210 at 7. [redacted] Nelson Decl. Ex. 63(a) at 321:19–323:9. [redacted] Lasinski Decl. Ex. B ¶¶ 182, 258.	Second, Plaintiffs fail to create a genuine dispute as to whether [redacted] The Microsoft documents cited by Plaintiffs do not controvert the stated fact. They were crafted *before* the [redacted] agreement was executed and do not reflect the reality after the agreement was actually executed. For example, the [redacted] it does not measure the value of books for training. Nelson Decl. Ex. 204 (MSFT_AI_MDL_001618003) at -057; *see* ECF 1707 (Usdan Decl.) ¶ 11. Plaintiffs' assertion that [redacted]
[redacted]	

[redacted]

Lasinski Decl. Ex. B ¶ 148 (citing Nelson Decl. Ex. 992 at -545). HarperCollins's CEO, Brian Murray, likewise stated that HarperCollins "w[as] able to reach a deal that we thought was good enough and worthwhile enough to take to authors and agents and give them the option to participate or not," and that "so far the vast majority of authors that we have spoken with have also opted in." Lasinski Decl. Ex. B ¶ 354 (citing Nelson Decl. Ex. 993).

[redacted] Lasinski Decl. Ex. B ¶ 149 (citing Nelson Decl. Ex. 994 at -373).

[redacted] Lasinski Decl. Ex. B ¶ 258; Lasinski Decl. Ex. C ¶ 128. HarperCollins, for its part, obtained and maintained at its own expense the approvals necessary to deliver the content, and its CEO explained that HarperCollins entered the agreement to demonstrate that there can be "a commercial market for licensing books for [redacted] is belied by the record. Mr. Usdan explained that [redacted] ECF 1707 (Usdan Decl.) ¶ 24(a). Additionally, the experiment was also [redacted] *Id.* ¶ 24(b). Thus, Plaintiffs' cited evidence reflecting hopes about books *before* the deal cannot contradict Microsoft's after-the-fact conclusion that it was a failure.

Third, Plaintiffs' cited evidence does not controvert that [redacted] Lasinski's opinion that the [redacted] is speculative and lacks foundation. ECF 1825 (Lasinski Decl.) Ex. B ¶ 148. Usdan's testimony establishes that Microsoft was [redacted] ECF 1707 (Usdan Decl.) ¶ 24(b).

Fourth, Plaintiffs fail to create a dispute as to whether [redacted] The only evidence that Plaintiffs cite to support the proposition that "the vast majority of authors that we have spoken with have also opted in" is from a podcast where HarperCollins's CEO was interviewed, which is unauthenticated and thus inadmissible hearsay as to Microsoft. Nelson Decl. Ex. 993.

training AI models." Lasinski Decl. Ex. B ¶¶ 258, 354; Lasinski Decl. Ex. C ¶ 128. [redacted] Lasinski Decl. Ex. B ¶ 256 (citing Nelson Decl. Ex. 197). [redacted] Lasinski Decl. Ex. B ¶ 258 (citing Nelson Decl. Ex. 63(b) at 79:23-25, 80:611); Lasinski Decl. Ex. C ¶ 128. [redacted] Zhao Decl. Ex. A ¶¶ 44-45; Lasinski Decl. Ex. B ¶ 258; Nelson Decl. Ex. 204 at -006, -016, -023, -057.	As explained by Mr. Usdan, the Microsoft employee "[redacted]." ECF 1707 (Usdan Decl.) ¶ 22. Plaintiffs' assertions concerning [redacted] does not negate the fact that [redacted] Fifth, Plaintiffs' cited evidence does not controvert that [redacted] Paragraph 28(g) [redacted] Plaintiffs' attempt to manufacture in dispute does not in fact dispute Paragraph 28(g).

Paragraph 29

29. After the release of the LLMs at issue in this case, Books Plaintiffs have continued to write and publish books. Tucker Decl. ¶¶ 14, 151; Hurst Decl. Ex. 128 (Wolicki Dep. Tr.) 26:20-27:3, 30:14-20; Ex. 87 (Picoult Dep. Tr.) 166:14-167:4; Ex. 108 (Shapiro Dep. Tr.) 46:1-12; Ex. 106 (Schiff Dep. Tr.) 20:16-21:6; Ex. 35 (Franzen Dep. Tr.) 61:14-18; Ex. 44 (Grisham Dep. Tr.) 178:21-25.	
Class Plaintiffs' Response: Undisputed that Class Plaintiffs have continued to write and publish books. **Disputed** to the extent Microsoft contends that continued publication demonstrates the absence of harm. As set forth in greater detail in response to Microsoft's Statement of Fact No. 25, Class Plaintiffs have identified market substitution and displacement of book sales, lost licensing revenues, obstacles to their ability to monetize their works, and reputational harm, none of which is negated by continued authorship.	**Microsoft's Reply: Undisputed** that after the release of the LLMs at issue, Books Plaintiffs have continued to write and publish books. The remainder of Books Plaintiffs' assertions constitute improper legal argument. To the extent Books Plaintiffs incorporate their response to Paragraph 25, Microsoft hereby incorporates its Reply to that paragraph as if set forth herein.

B. Lack of Volitional Conduct

Paragraph 30

30. Microsoft did not acquire the Books1 or Books2 datasets used by OpenAI as part of the training corpus in connection with the development of GPT 3 and 3.5. Hurst Decl. Ex. 120 (Tinter 30(b)(6) Dep. Tr.) 166:23-169:17, 13:6-22:18, 29:24-30:22; Ex. 120.A (Tinter Dep. Ex. 3 - Joint Development and Collaboration Agreement) 4-5 (§§1(f), 1(q)); Ex. 85 (Parakhin Dep. Tr.) 79:3-18, 15:18-21, 16:1-5, 18:7-12, 22:25-23:8, 23:21-25:6, 26:12-27:4; Ex. 118 (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 62:4-25, 296:12-18, 13:14-16, 14:14-20, 15:2-12, 16:18-19:10, 22:1-14, 22:25-23:15, 24:2-27:16; Ex. 96 (Radford Dep. Tr.) 37:17-41:11, 46:3-13, 117:4-20, 11:15-12:10, 13:22-14:6, 16:3-8, 17:13-19, 20:25-24:22; Ex. 295 (Mann Dep. Tr. (dated Mar. 3, 2025)) 58:14-60:4, 65:22-66:11, 13:19-20:5, 20:21-21:21; Ex. 48 (Hallacy Dep. Tr. (Vol. 1, dated Sept. 30, 2025)) 91:10-12, 92:5-18, 9:3-13, 9:23-10:9, 11:9-22, 29:16-25, 33:22-34:18; Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 95:10-23, 168:1-11, 169:15-170:17, 9:18-21, 15:2-25, 16:7-11, 17:2-18:7; Ex. 84 (Paino 30(b)(6) Dep. Tr.) 15:12-21:24, 22:12-26:6, 216:11-217:22, 12:18-25, 13:3-6, 13:22-14:7.	
Class Plaintiffs' Response: **Disputed.** Admissible evidence contradicts Microsoft's statement and shows that Microsoft did acquire the GPT-3 training data from OpenAI. As explained by Class Plaintiffs' expert Dr. Shan: [redacted] Shan Decl. Ex. A at ¶ 72(d) (citing Nelson Decl. Ex. 158 at -067).	**Microsoft's Reply:** Paragraph 30 is not genuinely disputed. Books Plaintiffs do not dispute that OpenAI employees acquired the underlying content to create the Books1 and Books2 datasets that were "used by OpenAI as part of the training corpus in connection with the development of GPT 3 and 3.5." BP SUMFs ¶¶ 138-160, 177-181, 192-243, 269-279. Plaintiffs' Opposition agrees that this fact is undisputed. *See* ECF 2268 BP Opp. 34. As the evidence cited by Microsoft in support of Paragraph 30 makes clear, Microsoft did not obtain the underlying works from the internet, select the contents of the datasets, or compile those works into the training datasets, and Plaintiffs' Response does not dispute this evidence. Plaintiffs seek to manufacture a different dispute by attributing a different meaning to the term "acquired" than how it is used in the fact as stated. Plaintiffs' assertion that "Microsoft did acquire the GPT-3 data *from OpenAI*" after

Benton Gaffney, Deputy General Counsel at OpenAI, testified that this provision entitled Microsoft to access the GPT-3 training datasets. Nelson Decl. Ex. 18 at 145:14–146:4. Microsoft accessed the GPT-3 training data in January 2021. Nelson Decl. Ex. 346. OpenAI also shared portions of the GPT-3 training data with Microsoft on at least one other occasion. Nelson Decl. Ex. 275 at -590–591. Evidence also shows that LibGen data was included in the GPT-3 training data. *See, e.g.*, Shan Decl. Ex. A at ¶ 51 ("A version of [Alec] Radford's LibGen dataset titled 'libGen1-dedup-clean' was used to train GPT-3.") (citing Nelson Decl. Ex. 62(b) at 50–52); Lasinski Decl. Ex. A at ¶ 67 (noting that Alex Paino "testified that OpenAI used both LibGen-1 and LibGen-2 data to train its GPT-3 model and variants of the data to train its GPT-3 model" (citing Nelson Decl. Ex. 43(a) at 234:20–235:22).	OpenAI had already used it to train GPT-3 does not controvert the stated fact. Plaintiffs' response is comprised only of additional facts that bear on that distinct assertion and do not controvert the stated fact. To the extent necessary, and out of an abundance of caution, Microsoft responds to these additional facts as follows: *First*, Plaintiffs' assertion that "[redacted]" is not supported by the cited evidence or the record for the reasons explained in Defs.' Resp. to BP SUMF ¶ 573. *Second*, to the extent Plaintiffs suggest Mr. Gaffney testified that the applicable JDCA provision granted Microsoft rights and/or indefinite access to the entirety of OpenAI's GPT-3 training datasets, that assertion mischaracterizes Mr. Gaffney's testimony and the terms of the governing JDCA amendment, as detailed in Defs.' Resp. to BP SUMF ¶ 574. Microsoft received only temporary access to a subset of the GPT-3 training datasets for the limited purpose of evaluating the GPT-3 model, and once the evaluations were completed, Microsoft was no longer entitled to access the GPT-3 training datasets. *See* Nelson Decl. Ex. 158 at -964 [redacted]); *see also* 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). Further, [redacted]

	[redacted] *See* Nelson Decl. Ex. 158 at -967. Exhibit E to the referenced agreement [redacted] *Id.* Third, to the extent Plaintiffs' assertion "Microsoft accessed the GPT-3 training data in January 2021" implies that Microsoft had access to the full GPT-3 training dataset, that implication is not supported by the cited evidence or the record. OpenAI only provided Microsoft with a subset of its GPT-3 training data in a tokenized format. *See* Blythe Opp. Decl. Ex. 39 (B. Norick Tr.) at 47:11 (explaining that OpenAI transferred "[a] fraction of pretraining data"); 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); Blythe Opp. Decl. Ex. 60 (X. Song 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."); Blythe Opp. Decl. Ex. 61 (J. Huynh 30(b)(6) Tr.) at 237:11–13 ("[T]he training datasets were tokenized, so not in natural language for a human to look at."). Finally, Books Plaintiffs' claims that "OpenAI also shared portions of the GPT-3 training data with Microsoft on at least one other occasion" and any implication that the "LibGen data was included" in the portions shared with Microsoft are not supported by Books Plaintiffs' own cited evidence. *See* Nelson Decl. Ex. 275 at -588 ([redacted]

	[redacted] This additional fact also does not controvert the stated fact.

Paragraph 31

31. Microsoft did not direct OpenAI to download the Books1 or Books2 datasets from Library Genesis ("LibGen") or any other so-called "shadow library." Hurst Decl. Ex. 120 (Tinter 30(b)(6) Dep. Tr.) 166:23-169:17; Ex. 120.A (Tinter Dep. Ex. 3 – Joint Development and Collaboration Agreement) 4-5 (§§1(f), 1(q)); Ex. 85 (Parakhin Dep. Tr.) 26:12-27:4, 79:3-18; Ex. 118 (Templeton 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 20, 2026)) 62:4-25, 296:12-18; Ex. 96 (Radford Dep. Tr.) 37:17-41:11, 46:3-13, 117:4-20; Ex. 295 (Mann Dep. Tr. (dated Mar. 3, 2025)) 58:14-60:4, 65:22-66:11, 13:19-20:5, 20:21-21:21; Ex. 48 (Hallacy Dep. Tr. (Vol. 1, dated Sept. 30, 2025)) 91:10-12, 92:5-18; Ex. 104 (Ryder 30(b)(6) Dep. Tr. (Vol. 1, dated Jan. 29, 2026)) 95:10-23, 168:1-11; 169:15-170:17; Ex. 84 (Paino 30(b)(6) Dep. Tr.) 15:12-21:24, 22:12-26:6, 216:11-217:22.	
Class Plaintiffs' Response: **Disputed** to the extent Microsoft contends that it was unaware that OpenAI's training data was sourced from Library Genesis or that it did not facilitate the acquisition of LibGen. First, admissible evidence demonstrates that Microsoft knew OpenAI's training data included LibGen. *See* Nelson Decl. Ex. 39 at 47:21–51:19; Nelson Decl. Ex. 59(a) at 292:2–17, 296:22–297:2, 299:19–23; Nelson Decl. Ex. 188 at -958; Nelson Decl. Ex. 189 at -964; Nelson Decl. Ex. 192 at -141; Nelson Decl. Ex. 661 at 28:12–29:10, 51:23–52:2; Nelson Decl. Ex. 696 at -654; Nelson Decl. Ex. 697; *see also* Class Plaintiffs' Statement of Facts Nos. 653–669. Second, admissible evidence further demonstrates that Microsoft made significant financial investments in OpenAI after learning that OpenAI used LibGen. *See* Nelson Decl. Ex. 2(a) at 219:15-220:13, 226:24–227:5; Nelson Decl. Ex. 145 at -347; Nelson Decl. Ex. 188 at -958; Nelson Decl. Ex. 203;	**Microsoft's Reply:** Paragraph 31 is not genuinely disputed. **Undisputed** that "Microsoft did not direct OpenAI to download the Books1 or Books2 datasets from Library Genesis or any other so-called 'shadow library.'" Plaintiffs' Response consists only of additional assertions that do not address or specifically controvert the fact stated in Paragraph 31. That does not make this fact disputed. In particular, to the extent that Plaintiffs' additional assertions purport to establish that Microsoft was "[]aware that OpenAI's training data was sourced from Library Genesis," such assertions do not controvert the stated fact and are immaterial to the issues before the Court; Microsoft's alleged "knowledge" is not an element of any claim or defense at issue in the summary judgment motions before the Court. To the extent that Plaintiffs also purport to establish that Microsoft "facilitate[d] the acquisition" of

Nelson Decl. Ex. 327 at -135; Nelson Decl. Ex. 330 at -573; Nelson Decl. Ex. 812 at 911–12; *see also* Class Plaintiffs' Statement of Facts Nos. 187–191. Third, despite having the contractual right to require OpenAI to retrain models using data Microsoft approved, Microsoft never did so. *See* Nelson Decl. Ex. 146 at -586, -590; Nelson Decl. Ex. 40 at 60:7–61:10; Nelson Decl. Ex. 60 at 126:20–22; Nelson Decl. Ex. 40 at 76:11–19, 76:20–77:2; *see also* Class Plaintiffs' Statement of Facts Nos. 688–700. Fourth, in August 2023, Microsoft started a scraping service for OpenAI dubbed "Project Mango," and included the LibGen domain libgen.ee in a list of "50k top domains to be crawled by MSFT," as set forth in greater detail in response to Microsoft's Statement of Fact No. 32. Lasinski Decl. Ex. A at ¶ 91; Nelson Decl. Ex. 867 at 75:11–13, 79:8–12, 102:4–103:2.	the Books1 and Books2 datasets through such knowledge, those statements too do not controvert the stated fact and are immaterial for the same reasons. All such statements should therefore be disregarded. To the extent necessary, and out of an abundance of caution, Microsoft responds to Plaintiffs' additional assertions as follows: *First*, Plaintiffs' purported facts "demonstrat[ing] that Microsoft knew OpenAI's training data included LibGen" or that Microsoft made investments in OpenAI "after learning that OpenAI used LibGen" are contradicted by Plaintiffs' own statements of "undisputed" material fact. *See* BP Counter-SUMF ¶ 23 (asserting that Microsoft witnesses testified that they were unaware "as to both what LibGen is and whether OpenAI used it"). *Second*, Plaintiffs' cited evidence does not state, suggest, or otherwise support that Microsoft "approved" of OpenAI's use of the LibGen datasets. The retraining clause in the agreement cited as Nelson Decl. Ex. 146 was only in place beginning in January 2023, which Plaintiffs do not dispute was years *after* OpenAI employees downloaded the data from LibGen that comprises the Books1 and Books2 datasets and used it to train GPT-3 and GPT-3.5. BP SUMF ¶¶ 160, 193, 210, 217, 220, 227, 233, 521, 526. Microsoft thus could not have exercised any retraining right with respect to the GPT models trained on Books1 and Books2. In any case, the 2023 JDCA provides that Microsoft could require OpenAI to retrain models using data Microsoft approved *only* if Microsoft "has concerns regarding the use of any particular data to train a Model." Nelson Decl. Ex. 146 at -590. Plaintiffs' cited deposition testimony does not support that Microsoft became aware of any information that would raise "concerns" regarding the datasets used to

	train any OpenAI model. The evidence cited in Paragraph 31 contradicts any such assumption. Hurst Decl. Ex. 120 (J. Tinter 30(b)(6) Tr.) at 170:18-25 ("Q. And why didn't Microsoft ever exercise its contractual right to require OpenAI to promptly retrain a model using solely data approved by Microsoft or its affiliate?" "A. … We never became aware of – we never developed a concern about the commercialization ability of a model."). The 2023 JDCA further provides that, "[s]ubject to the foregoing, OpCo and OpenAI may train all Models using data selected solely by OpCo and OpenAI[.]" Nelson Decl. Ex. 146 at -590; *see also* Hurst Decl. Ex. 120 (J. Tinter 30(b)(6) Tr.) at 167:7-11 ("I think to understand the full context here, ultimately, at the end of the day, the decision on what data was used was 100 percent OpenAI's decision, right. They made that decision."). *Third*, while the Project Mango list of domains for Microsoft to crawl for OpenAI included the LibGen domain libgen.ee, the cited evidence does not support that Microsoft crawled content from the libgen.ee domain or that such content was available to crawl. *See* Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 66:5-69:5; Microsoft's Reply to Paragraph 32. Further, Plaintiffs' expert Shan did not identify any content from these domains in OpenAI's GPTBot dataset purportedly crawled by Microsoft. *See* ECF 1814-1 (Shan Decl.) Ex. A, ¶ 137; ECF 1814 (Shan Decl.) Ex. A, Appendix L.

Paragraph 32

32. To the extent Microsoft acquired data to be used by OpenAI in connection with the development of LLMs, Microsoft did so by using ordinary software robots that crawled content from URLs across the public internet; Microsoft's web robots were not capable of utilizing the Torrent protocol implemented by so-called "shadow libraries." Gaggara Decl. ¶¶ 20-21; Hurst Decl. Ex. 38 (Gaggara 30(b)(6) Dep. Tr. (dated Aug. 20, 2025)) 67:1-20, 143:13-144:6, 8:9-19, 9:25-13:17.

Class Plaintiffs' Response: Class Plaintiffs object to the phrase "ordinary software robots" as vague and ambiguous. Subject to and without waiving this objection, **Undisputed** that Microsoft's web crawlers did not themselves implement the Torrent protocol. **Disputed** to the extent Microsoft contends that its crawling was confined to ordinary public-internet content or that the limits of its crawlers establish that Microsoft did not acquire or host data sourced from shadow libraries. Admissible evidence contradicts Microsoft's statement. According to Class Plaintiffs' expert Dr. Shan: [redacted] Jamie Huynh testified that he understood each of these links to be "books, individual books." Shan Decl. Ex. A at ¶ 130 (citing Nelson Decl. Ex. 171 at slide 2 & Nelson Decl. Ex. 873 at 85:11–17). Further, in connection with Project Mango, the list of "50k top domains to be crawled by MSFT" from the Bing Index included the LibGen domain, Libgen.ee, among others.	**Microsoft's Reply:** Paragraph 32 is not genuinely disputed. **Undisputed** that "[t]o the extent Microsoft acquired data to be used by OpenAI in connection with the development of LLMs, Microsoft did so by using its [web crawler] that crawled content from URLs across the [] internet" and that "Microsoft's web [crawler] [was] not capable of utilizing the Torrent protocol." Plaintiffs' assertions that Microsoft's crawling was not confined to "ordinary public-internet content" and included "data sourced from shadow libraries" do not controvert the stated fact and are not supported by the cited evidence or the record. *First*, Plaintiffs' assertion that [redacted] does not controvert the stated fact and is not supported by the cited evidence. Nelson Decl. Ex. 171 is an internal Microsoft presentation dated October 2023 describing prospective "goals" for a "workstream" that was explicitly separate from "Bing Crawl." *Id.* at Slides 2-3; *see also* Hurst Decl. Ex. 359 (J. Usdan 30(b)(6) Tr.) at 52:25-53:2, 147:13-149:10. Plaintiffs cite no evidence that this presentation had anything to do with Project Mango, which is the subject of the stated fact.

Lasinski Decl. Ex. A at ¶ 91. As part of Project Mango, OpenAI paid Microsoft "to crawl the web and then transfer the material from its Azure storage accounts into Azure storage accounts associated with OpenAI." Shan Decl. Ex. A at ¶ 135; Nelson Decl. Ex. 57(a) at 336:20–344:9. Evidence also shows that Microsoft's infrastructure held torrent-sourced LibGen data. *See* Choffnes Decl. Ex. A at ¶ 136 (citing Nelson Decl. Ex. 264, a work chat stating that "raw libgen (fiction & nonfiction)" was copied from Google storage to Microsoft Azure storage). That data was obtained by torrenting rather than by any ordinary "web robot". From Fall 2019 to January 2020, OpenAI employees Ben Mann and Chris Hallacy used special torrenting software and torrented both the fiction and nonfiction components of Library Genesis, including all 2,176 LibGen torrents. Nelson Decl. Ex. 33 at 145:2–11; Nelson Decl. Ex. 803 at -998; Shan Decl. Ex. A at ¶¶ 46, 52–60. Data acquisition likewise extended beyond ordinary crawling in other respects: OpenAI President Greg Brockman wrote and ran his own "private scraping cluster" to re-download PDFs that Common Crawl had truncated. Nelson Decl. Ex. 8 at 213:25–214:24; Shan Decl. Ex. A at ¶ 122.	The Shan evidence that Plaintiffs cite is based on a slide entitled "Books" that lists "Library Genesis" (among public domain libraries and archives such as "Google Books") with a note "300K in index, about another 300K is blocked." Nelson Decl. Ex. 171 at Slide 17. The cited evidence does not suggest or support that Microsoft acquired and provided to OpenAI content from the "LibGen URLs" discussed in the document cited as Nelson Decl. Ex. 171. Rather, [redacted] Hurst Decl. Ex. 359 (J. Usdan 30(b)(6) Tr.) at 52:25-53:2; *see also* Hurst Decl. Ex. 350 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 23:11-25:4. Nor does the cited evidence support Plaintiffs' assertion that "each of these links" was an "individual book[]." When presented with Slide 17 from Nelson Decl. Ex. 171, Mr. Huynh testified, "It is unclear what the metric is here. It could be books. URLs could map directly to a number of books." Nelson Decl. Ex. 873 (J. Hyunh 30(b)(6) Tr.) at 85:7-9. When further questioned about the meaning of "blocked" in this context, Mr. Huynh testified, "I'm not a crawler. I -- I did not work on the crawler itself" and "I'm unsure. I would – I would have to speculate, be speculating." *Id.* at 86:2-18. Ms. Gaggara, Microsoft's corporate designee concerning the LibGen URLs contained in the Bing Index, confirmed that the mere existence of a URL in the Bing Index does not mean that there is meaningful content available on that webpage to be indexed or that Microsoft's Bingbot crawler can access the contents that may exist on a given webpage. *See* ECF 1705 (Gaggara Decl.) ¶¶ 20-22; Blythe Opp. Decl. Ex. 63 (P. Gaggara 30(b)(6) Tr. (Aug. 20, 2025)) at 28:15–31:8; *see also* Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 68:20-69:5.

	Second, Plaintiffs' suggestion that Microsoft crawled content from the "LibGen domain" or "Libgen.ee" and provided such content to OpenAI in connection with Project Mango is unfounded speculation that is not supported by the cited evidence or the record. Plaintiffs do not dispute that Microsoft used the existing codebase for Bingbot, which was not capable of utilizing the Torrent protocol. Plaintiffs' cited evidence establishes only that the LibGen domains were in the list of domains to be crawled by Microsoft, but 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* Blythe Opp. Decl. Ex. 85 (F. Canel 30(b)(6) Tr.) at 66:5-69:5. Indeed, Plaintiffs' expert Shan did not identify any content from these domains in OpenAI's GPTBot dataset purportedly crawled by Microsoft. *See* ECF 1814-1 (Shan Decl.) Ex. A, ¶ 137; ECF 1814 (Shan Decl.) Ex. A, Appendix L. *Third*, Plaintiffs' assertion that "Microsoft's infrastructure held torrent-sourced LibGen data" that OpenAI's employees acquired independent of Microsoft does not address or controvert the facts stated in Paragraph 31, which specifically concerns *Microsoft's* acquisition of data for OpenAI using *Microsoft's web crawler* that, as Plaintiffs concede, *cannot access or collect content only available through torrenting*. Plaintiffs' cited evidence shows only that *OpenAI* stored certain content in Microsoft's Azure servers, which Microsoft had no visibility into. *See* Simpson Decl. Ex. 335 (L. Melton 30(b)(6) Tr.) at 65:16-66:19.

II. COUNT II: VICARIOUS COPYRIGHT INFRINGEMENT

A. No underlying direct infringement

Paragraph 33

33. Microsoft incorporates Undisputed Facts 1-32 as though fully restated herein.	
Class Plaintiffs' Response: **Disputed.** This is not an appropriate undisputed fact. To the extent that the Court considers this "fact," Class Plaintiffs incorporate by reference their responses to Microsoft's Undisputed Facts Nos. 1-32 as though fully restated herein.	**Microsoft's Reply:** Microsoft disagrees that this is inappropriate and incorporates its Replies to Undisputed Facts 1-32 as though fully restated herein.

III. COUNT III: CONTRIBUTORY COPYRIGHT INFRINGEMENT

A. No underlying direct infringement

Paragraph 34

34. Microsoft incorporates Undisputed Facts 1-33 as though fully restated herein.	
Class Plaintiffs' Response: **Disputed.** This is not an appropriate undisputed fact. To the extent that the Court considers this "fact," Class Plaintiffs incorporate by reference their responses to Microsoft's Undisputed Facts Nos. 1-33 as though fully restated herein.	**Microsoft's Reply:** Microsoft disagrees that this is inappropriate and incorporates its Replies to Undisputed Facts 1-33 as though fully restated herein.

Dated: November 20, 2026

Respectfully submitted,

/s/ Annette L. Hurst

ORRICK, HERRINGTON & SUTCLIFFE LLP
Annette L. Hurst (admitted pro hac vice)
Daniel Justice (admitted pro hac vice)
Robert Scott Shwarts
Catherine Lui (admitted pro hac vice)
Amanda H. Schwartz (admitted pro hac vice)
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 Berkeley 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

CERTIFICATE OF SERVICE

I hereby certify that on November 20, 2026, I caused the foregoing document to be electronically filed with the Clerk of the United States District Court for the Southern District of New York using the CM/ECF system, which shall send electronic notification to all counsel of record.

/s/ Annette L. Hurst