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 Case No. 1:23-cv-10211	Case No. 1:25-md-3143-SHS-OTW **REDACTED – PUBLIC VERSION**

Defendant Microsoft Corporation's Memorandum of Law In Support Of Motion For Summary Judgment In Books Plaintiffs' Consolidated Cases

TABLE OF CONTENTS

Page

TABLE OF AUTHORITIES ii
TABLE OF CITATION CONVENTIONS v
ORIENTATION 1
INTRODUCTION 1
STATEMENT OF FACTS 4
OpenAI and Microsoft partner to develop Large Language Models. 4
LLMs are trained on a large and varied corpus of data, including books, to create a general purpose technology. 6
Book authors claim that training an LLM infringes their copyrights, but struggle in discovery to substantiate their feared harms. 8
ARGUMENT 9
I. Training An LLM Is Fair Use As A Matter Of Law. 9
A. The purpose and character of the use weighs overwhelmingly in favor of fair use (Factor one). 10
1. The use of books to train an LLM is profoundly transformative. 10
2. Plaintiffs' arguments based on "shadow libraries," "memorization," and the Bing Index do not weigh against the transformativeness of LLM training. 17
B. The nature of the underlying works and the amount of them used carry little weight because of the transformativeness of the use (Factors two and three). 24
C. Training an LLM does not cause substantial market harm in any cognizable market (Factor four). 25
1. Training an LLM does not substitute for books in their traditional markets. 25
2. The Court should reject Plaintiffs' arguments regarding a new AI training market and theory of market dilution from non-infringing works. 26
CONCLUSION 33
CERTIFICATE OF WORD COUNT COMPLIANCE 36

TABLE OF AUTHORITIES

Page(s)

Cases

Am. Geophysical Union v. Texaco Inc.,
60 F.3d 913 (2d Cir. 1994) 14, 16, 21, 25, 28

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith,
598 U.S. 508 (2023) 10, 18, 30, 31

Authors Guild, Inc. v. HathiTrust,
755 F.3d 87 (2d Cir. 2014) 10, 13, 24, 25, 27

Authors Guild v. Google, Inc.,
804 F.3d 202 (2d Cir. 2015) 10, 11, 12, 13, 14, 16, 18, 24, 25

Baker v. Selden,
101 U.S. 99 (1879) 30

Bartz v. Anthropic PBC,
787 F. Supp. 3d 1007 (N.D. Cal. 2025) 11, 15, 17, 21, 27

Blanch v. Koons,
467 F.3d 244 (2d Cir. 2006) 16

Cambridge Univ. Press v. Patton,
769 F.3d 1232 (11th Cir. 2014) 28

Campbell v. Acuff-Rose Music, Inc.,
510 U.S. 569 (1994) 10, 19

Cartoon Network LP, LLLP v. CSC Holdings, Inc.,
536 F.3d 121 (2d Cir. 2008) 18

Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.,
150 F.3d 132 (2d Cir. 1998) 28, 31

Cengage Learning, Inc. et al. v. Doe et al.,
No. 23-cv-8136-CRM, ECF No. 36 (S.D.N.Y. Sept. 24, 2024) 20

Consumers Union of U.S., Inc. v. Gen. Signal Corp.,
724 F.2d 1044 (2d Cir. 1983) 16

Cox Commc'ns, Inc. v. Sony Music Ent.,
146 S. Ct. 959 (2026) 9

Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.,
499 U.S. 340 (1991) 31

Fox Broad. Co. v. Dish Network L.L.C.,
747 F.3d 1060 (9th Cir. 2014) 18

Google LLC v. Oracle Am., Inc.,
593 U.S. 1 (2021) 10, 13, 14, 18, 19, 26, 27

Hachette Book Grp., Inc. v. Internet Archive,
115 F.4th 163 (2d Cir. 2024) 26

Harper & Row Publishers, Inc. v. Nation Enters.,
471 U.S. 539 (1985) 19

Infinity Broad. Corp. v. Kirkwood,
150 F.3d 104 (2d Cir. 1998) 25

Kadrey v. Meta Platforms, Inc.,
788 F. Supp. 3d 1026 (N.D. Cal. 2025) 11, 13, 18, 20, 23, 26, 27, 30, 31

Mattel, Inc. v. Goldberger Doll Mfg. Co.,
365 F.3d 133 (2d Cir. 2004) 30

McDonald v. West,
138 F. Supp. 3d 448 (S.D.N.Y. 2015) 15

NXIVM Corp. v. Ross Inst.,
364 F.3d 471 (2d Cir. 2004) 16, 19

Perfect 10, Inc. v. Amazon.com, Inc.,
508 F.3d 1146 (9th Cir. 2007) 23

Sega Enters. Ltd. v. Accolade, Inc.,
977 F.2d 1510 (9th Cir. 1992) 13, 15, 31

Sony Comp. Ent., Inc. v. Connectix Corp.,
203 F.3d 596 (9th Cir. 2000) 12, 15, 24, 31

Superior Form Builders, Inc. v. Dan Chase Taxidermy Supply Co.,
74 F.3d 488 (4th Cir. 1996) 30

Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.,
756 F.3d 73 (2d Cir. 2014) 28

Whyte Monkee Prods., LLC v. Netflix, Inc.,
174 F.4th 761 (10th Cir. 2026) 28

Statutes

17 U.S.C. § 102(a) 28

17 U.S.C. § 102(b) 15, 30

17 U.S.C. § 107 ..... 10, 16, 24, 25

Other Authorities

Betty Alexandra Toole, *Poetical Science*, 15 The Byron J. 55 (1987) 1

4 *Nimmer on Copyright* § 13.05[A][4] 27

4 *Nimmer on Copyright* § 13F.14[B] 21

Patry on Fair Use § 6.13 26

Pierre N. Leval, *Toward a Fair Use Standard*, 103 Harv L. Rev. 1105 (1990) 11, 20

TABLE OF CITATION CONVENTIONS

This brief uses the following citation conventions for record materials filed in support of Microsoft's Motion for Summary Judgment.

Citation	**Document**
Hurst Ex.	Declaration of Annette L. Hurst in Support of Microsoft's Motions for Summary Judgment
Lafferty ¶	Declaration of John D. Lafferty, Ph.D. in Support of Microsoft's Motions for Summary Judgment
SUMF ¶	Rule 56.1 Statement of Undisputed Material Facts in Support of Microsoft's Books Motion for Summary Judgment
Tucker ¶	Declaration of Catherine Tucker, Ph.D. in Support of Microsoft's Books Motion for Summary Judgment

All statutory cites are to Title 17 unless otherwise indicated. All ECF cites are to MDL No. 1:25-md-3143-SHS-OTW unless otherwise indicated.

ORIENTATION

Microsoft recommends that this brief be reviewed first, followed by its brief in the News cases.

INTRODUCTION

The famous poet Lord Byron and his wife Annabella, an educational reformer, thought that literature and science were at odds. But their daughter Ada Lovelace, a brilliant mathematician and history's first computer scientist, thought differently. Lovelace made the case for "poetical science," and when she looked at primitive designs for a computer, she saw not a machine for long arithmetic, but an engine of progress that could wield any complex system humans could devise.[1] Ada was right. What better proof of poetical science than the advance at issue in this case: the Large Language Model, a computer system that harnesses the deep structure of language and knowledge to give people a profound tool for speech, communication, and learning.

Copyright law, too, is an engine of progress. It is founded in the Constitution's charge that Congress pass laws "[t]o promote the Progress of Science," which at the time referred to the expansion of knowledge and learning. In keeping with this aim, the Copyright Act creates qualified rights in authors to exploit their own expressive works, while safeguarding the public's rights to use and build upon our collective intellectual potential. This balance has worked, fostering new technologies for speech and communication that lead to still more creation and innovation. The law accomplishes this principally by protecting the public's right to fair use. Fair use is why search engines are legal, why viewers can record their favorite shows to watch later, why

[1] *See* Betty Alexandra Toole, *Poetical Science*, 15 The Byron J. 55 (1987).

software developers are free to make apps that work on various computers and smartphones. Rights owners claimed that all these transformative technologies should be barred on the theory that any use of their works required payment. Each time the law said otherwise, the rights owners adapted, and our world was richer for it.

Here we are again. The book authors who bring this case assert an unqualified right to prohibit the use of their works to train an LLM. But as two courts have already held, and as undisputed evidence here confirms, training an LLM is an exceedingly transformative way to use written works. What results from this use is an engine for communication, speech, and learning so foundational that it powers countless new tools for humans to wield, while making existing ones even more useful. As for the effect of LLM development on the market for books, Plaintiffs began this case speculating that LLMs would destroy their very livelihoods. Years of discovery later, the record is clear: Neither LLM training nor the use of LLMs in products substitutes for copyrighted books, in purpose or in practice.

Plaintiffs do not seriously contest that LLMs are transformative technology. They instead seek to deflect. Their principal tack is to point to the "pirate" website from which OpenAI allegedly sourced some books. They claim that the provenance of the training data—and the fact that *someone else* previously wronged them by infringement through that website—weighs against a finding of fair use. But this theory is unmoored from the law and logic of fair use. It is a given that any use for which a defendant asserts a fair use defense is unauthorized. What matters to the legality of the use is its purpose and character, and whether the use is a substitute for a use in a market a plaintiff is exclusively entitled to exploit. How exactly a defendant came upon the copy of the work is incidental to any transformative use of the work.

This Court should hold that the use of copyrighted works to train a general-purpose LLM is fair use as a matter of law. The fair use factors—each of which is informed by the overwhelming transformativeness of the use—dramatically favor that result. So, ultimately, does the public interest in knowledge and learning. Already, LLM-based tools are used for a host of expressive and knowledge-based uses that copyright law cherishes. They tailor education to the individual needs of young students, help pro bono attorneys analyze evidence and develop legal theories more efficiently, assist scientists researching new compounds, lead to medical diagnostic breakthroughs, and perform countless more functions that help us work, live, and learn. Indeed, even amidst this dispute, most of those involved in this case—Plaintiffs and Defendants, lawyers and experts—are already using LLM-based tools regularly.

As is always the case with technological breakthroughs of this magnitude, LLMs have critics and detractors who raise a variety of concerns about societal risks. And doubtless any technology this powerful requires care in development and deployment, which Microsoft continues to demonstrate by establishing industry norms and seeking to minimize any harmful effects. But a copyright case is not the place to litigate every policy implication that attends LLMs. Training an LLM is fair use as a matter of law, fair use is not copyright infringement, and that conclusion compels summary judgment on all of Plaintiffs' claims. The motion should be granted.

STATEMENT OF FACTS

OpenAI and Microsoft partner to develop Large Language Models.

The notion of computers that possess human-level intelligence has fascinated computer scientists for generations. Lafferty ¶ 36. But not until recent advances in the subfield of machine learning did Artificial Intelligence start to earn the title. These newer machine learning techniques are premised on a key insight: that even the most complex real-world systems like language or music have certain intrinsic patterns that are reflected in, and can be "learned" from, the vast collective store of data created by humans. Rather than impose fixed programming rules, these techniques seek to enable a computer system to identify and model patterns in data from real-world contexts. Lafferty ¶ 36. The process is called training. When complete, the resulting AI model has the ability to generate new content based on the learned statistics embedded in its math. Lafferty ¶ 46.

Early efforts to train a generative AI model based on natural language faced significant challenges. One problem was that existing approaches to training proceeded through text one word at a time, preventing them from deriving larger context. By the time a machine reached the end of a long passage, the beginning had often faded, leaving it unable to connect ideas that sat far apart. Lafferty ¶ 124. Another challenge was (and still is) that machine learning requires enormous amounts of data, and the more complex and varied the domain, the greater the volume and diversity of data needed. SUMF ¶ 2. Mapping the universe of written language requires a representative sample of that universe.

In 2017, a key breakthrough solved the first of these challenges. The innovation, known as the "transformer," is a model architecture capable of examining entire

sequences of words simultaneously rather than marching through them one at a time, thereby allowing the computer to capture far more (and far more nuanced) semantic and informational relationships between words. Lafferty ¶¶ 100, 124-25. The transformer thus made it possible to create a general-purpose model with so vast a network of mathematical parameters that it could generate a new, context-appropriate output in response to near any prompt. Lafferty ¶¶ 90, 100, 124-25. But there remained the significant issue of obtaining an astronomical amount of training data, not to mention securing the computing infrastructure to process it all. And whether the transformer architecture would ever yield a high-performing model was speculative, even as researchers raced to find out.

OpenAI was home to many of those researchers. In 2018, OpenAI first applied the transformer architecture to an untested model designed to generate text, developing its Generative Pre-trained Transformer, or "GPT." Hurst Ex. 226. This was proof of concept. OpenAI could amass the required volume of data by scraping everything publicly available on the internet, and had the engineering chops to use it. But OpenAI lacked the supercomputing power and related infrastructure—the "compute"—necessary to build, train, and operate ever more complex models. Hurst Ex. 3, 197:18-25, 198:4-9; Hurst Ex. 75, 22:20-24:13, 30:8-14, 91:13-20.

In stepped Microsoft. As a company built on developing new technologies and getting them to consumers, Microsoft was undaunted by the reality that OpenAI was still just "a research lab." Hurst Ex. 81, 250:20-251:24. It wanted to "sponsor anyone with ambition to compete at the frontier of AI." *Id.* at 250:20-25. As a "platform company," *id.* at 251:6-11, Microsoft knew how to scale and deploy new tools. And it could build the compute to make it happen. In 2019, Microsoft agreed to partner with

OpenAI to bring OpenAI's GPT models to the public. Hurst Ex. 120, 27:9-29:10. Microsoft provided OpenAI with both the substantial capital investment and the supercomputing infrastructure that OpenAI needed to train and deploy successive generations of GPT models at scale. Hurst Ex. 120, 27:9-29:4; Hurst Ex. 126, 16:2-20. Microsoft's vision was to responsibly develop "AI that is then going to be diffused broadly through the economy and society," Hurst Ex. 81, 249:4-13, as part of its longstanding mission to empower every person and every organization on the planet to achieve more.

LLMs are trained on a large and varied corpus of data, including books, to create a general purpose technology.

An LLM is a type of neural network—a computational model for processing information. This information (converted into a numerical representation) is passed through many layers of interconnected mathematical units; each layer involves a computation, based on the model's parameters, that is passed to the next layer until the model generates an output. SUMF ¶¶ 4.b-d. Training is the iterative process by which the model gradually adjusts the parameters that allow it to receive an input and generate a desirable output. *See* SUMF ¶ 4.d-e.

The first step is to acquire training data and assemble a training set. SUMF ¶ 4.a. As explained, because the point of an LLM is to derive semantic patterns reflecting as much of human language as possible, the LLM must be trained on data that is correspondingly expansive. SUMF ¶¶ 5, 7, 9. It is hard to overstate just how much data is used to train an LLM. For technical reasons explained in a moment, the relevant unit for training data is the token—a short word or portion of a longer word. Publicly disclosed training sizes for frontier LLM models range from 2 trillion tokens in 2023 to

roughly 18 trillion by late 2024. Lafferty ¶ 93; *see* SUMF ¶¶ 5-6. For scale: There are 100 billion stars in the Milky Way galaxy—it would take 180 such galaxies to count to 18 trillion.

As for variety, a high-performing, general-purpose LLM must respond to all manner of prompts, and so must be conversant in too many ways and on too many topics to name. SUMF ¶ 9. The only place to get data that is sufficiently voluminous and varied is the internet. SUMF ¶ 8. Model developers therefore "scrape" the internet—a longstanding practice deployed for a variety of purposes—trawling to amass all manner of data for use in model training. SUMF ¶ 4.a; *see* Tucker ¶¶ 117-18 (describing web crawling and scraping).

Once engineers have acquired and selected an optimal training corpus, the data is converted from text to numbers. Words are split into "tokens" that are then assigned a unique integer. SUMF ¶ 4.b. After text is tokenized, training moves to an iterative process of computing the relationships among all tokens, resulting in "embeddings" that place tokens in high-dimensional, mathematical space relative to all other tokens. SUMF ¶ 4.c. With enough text to analyze, the model is able to capture even very subtle relationships and distinctions between words, ideas, and concepts. SUMF ¶¶ 9, 12, 13, 19-20.

The fundamental objective of the training process is *generalization*. SUMF ¶¶ 4.e, 9. Generalization is what allows the model to generate coherent, contextually appropriate text about new subjects—that is, to meet unseen prompts with novel expression. SUMF ¶ 9. The training process is designed to create a vast mathematical operation with parameters representing the deeper structure of language, along with embedded relationships between facts, ideas, concepts, and so forth that

exist in the world. Lafferty ¶¶ 145, 155-56. And that explains why considerable volume and variety of data is so vital to a model's performance. Generalization requires discerning the underlying patterns and relationships revealed *across* works; it is an emergent property of the training dataset as a whole that is improved by broader coverage. SUMF ¶ 10. Models therefore perform best when they are trained on a wide variety of sources, ranging from books and web articles to social media posts and software code. SUMF ¶¶ 7-9.

Because it is a giant mathematical function that takes an input and generates an output, SUMF ¶¶ 1, 12, an LLM is not itself a finished product or tool. You cannot use an LLM any more than you can drive an engine. To power a chatbot, software coding tool, research assistant, AI agent, etc., the LLM must be incorporated into a stack of other technologies, which collectively dictate the product or tool's performance as a whole. Microsoft, OpenAI, and other innovators have already deployed LLMs in a first wave of different products and tools, which end-users have put to countless and growing uses.

Book authors claim that training an LLM infringes their copyrights, but struggle in discovery to substantiate their feared harms.

The Authors Guild and various named fiction and nonfiction authors brought this putative class action just ten months after OpenAI first launched its ChatGPT chatbot to the public in November 2022. ECF 1 (No. 1:23-cv-08292). After adding Microsoft as a Defendant, Plaintiffs later filed a Consolidated Class Action Complaint, ECF 183 ("Compl."). They alleged that "LLMs endanger authors' ability to make a living" and that the use of their works to train LLMs makes those works "into engines of Plaintiffs' own destruction." Compl. ¶ 3. They claimed that the OpenAI-trained GPT models at

issue were trained "so that the models could memorize, mimic, and paraphrase [] expression." Compl. ¶ 7; ECF 707 (order addressing models at issue).

Books Plaintiffs' allegations against Microsoft were narrow. The only LLMs addressed in the Complaint are OpenAI's models. And the Complaint contains no allegations concerning the operation, use, or output of any Microsoft product. Plaintiffs asserted claims against Microsoft for alleged direct copyright infringement by "reproducing the works owned by Plaintiffs … to create and develop datasets used to train … artificial intelligence models," Compl. ¶ 311, and for purported vicarious and contributory liability for OpenAI's training of its GPT models.[2]

Yet through years of discovery, Plaintiffs found nothing to substantiate the harms they feared. They are selling just as many books as they would have had ChatGPT and Copilot never been released. SUMF ¶¶ 22-24. They found virtually no examples of real-world outputs matching their works. SUMF ¶ 15. And while Books Plaintiffs have suggested that they will be harmed if others pick up LLM-based products and write competing books—referred to as market "dilution"—there is no evidence of any such dilution of the market for well-known authors' books. SUMF ¶ 26.

ARGUMENT

I. Training An LLM Is Fair Use As A Matter Of Law.

The fair use doctrine calls for "a multifaceted assessment of [a] crucial question: how to define the boundary limit of the original author's exclusive rights in order to best

[2] Plaintiffs' contributory liability claim, which is foreclosed by *Cox Communications, Inc. v. Sony Music Entertainment*, 146 S. Ct. 959 (2026), is also the subject of a motion for judgment on the pleadings Microsoft has filed contemporaneously with this motion. A finding that LLM training is fair use, as argued in this motion, would independently foreclose the contributory infringement claim, too.

serve the overall objectives of the copyright law[.]" *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015). The factors courts consider[3] are designed to weigh "incentives to create against the costs of restrictions on copying." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 526 (2023). The fair use inquiry is a mixed question of law and fact. Where material factual questions are undisputed, it is for the court to consider and balance the fair use factors, and to resolve the "ultimate question" as a matter of law. *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 24-25 (2021).

That resolution is appropriate here. Undisputed record evidence establishes that training an LLM is a phenomenally transformative way to use copyrighted works. The result is a foundational technology for speech, communication, and learning that will power countless new technologies, driving innovation and competition. And neither the acquisition and use of books to train LLMs nor LLMs themselves substitute for books in any traditional or likely to be developed book market. LLM training is fair use, and the Court should grant summary judgment on all claims.

A. The purpose and character of the use weighs overwhelmingly in favor of fair use (Factor one).

1. The use of books to train an LLM is profoundly transformative.

At the "heart" of the fair use doctrine is the question whether use of a copyrighted work is "transformative." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994). "A use that has a further purpose or different character" than the original, *Warhol*, 598 U.S. at 528-29 (citation omitted), or that "serves a new and different function," *Authors*

[3] Courts weigh a non-exclusive set of factors, including (1) the purpose and character of the use, (2) the nature of the copyrighted work, (3) the amount and substantiality of the portion used, and (4) the effect of the use upon the market. § 107.

Guild, Inc. v. HathiTrust, 755 F.3d 87, 96 (2d Cir. 2014), is transformative. "The more the appropriator is using the copied material for new, transformative purposes, the more it serves copyright's goal of enriching public knowledge[.]" *Authors Guild*, 804 F.3d at 214.

a. Training is transformative. As two courts have held, the use of copyrighted works to train an LLM is "highly" transformative, and there is no "serious question" about it. *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1044-45 (N.D. Cal. 2025); *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1019-22 (N.D. Cal. 2025) (training is "spectacularly" and "exceedingly" transformative).

Books are created "to be read." *Kadrey*, 788 F. Supp. 3d at 1044; SUMF ¶ 18. John Grisham's *The Street Lawyer* is sold in bookstores and downloaded on Kindles for entertainment on the beach or the bus. Stacy Schiff's *Cleopatra* offers readers an edifying narrative on an important historical figure. OpenAI's acquisition and use of *The Street Lawyer* or *Cleopatra* is not for purposes of reading those books, nor enabling others to read those books. The text of the books—along with trillions more tokens of text—is used for a technological purpose: to create an artificial intelligence model that can generate natural language responses to user prompts. That is a new and different way to use a book.

Undisputed evidence confirms the different character and purpose of OpenAI's use of copyrighted works to train an LLM.

Start with the transformative character. A use that "employ[s] the [copyrighted content] in a different manner … from the original" is transformative in character—as, for example, where copyrighted content is used as "raw material" for something "new." Pierre N. Leval, *Toward a Fair Use Standard*, 103 Harv L. Rev. 1105, 1111 (1990).

Another example is reverse engineering copyrighted computer code—the code is copied to study the (unprotected) logic of how it works, not to use the program itself. *Cf. Sony Comp. Ent., Inc. v. Connectix Corp.*, 203 F.3d 596, 608 (9th Cir. 2000) (holding that reverse engineering was fair use).

Books are used in LLM training in a different manner than readers use them. Coincidentally, the neural network architecture used to train LLMs is called a "transformer." Lafferty ¶ 124. And its operation well earns the moniker. The transformer breaks ordinary readable text into "tokens" and assigns each token a unique integer. SUMF ¶ 4.b. The tokens (combined with trillions of others) then influence embeddings in high-dimensional space—numerical vectors calibrated relative to one another to reflect relationships gleaned from the entirety of the training set. SUMF ¶ 4.c. The resulting model is a vast "mathematical operation" comprised of hundreds of billions of "parameters" used to predict the next token in a string of text. Lafferty ¶¶ 56-66, 137-41. That technology goes vastly beyond a book, a library, or, say, a searchable digital database of books. *Cf. Authors Guild*, 804 F.3d at 225 (holding that searchable book database is transformative fair use).

As for the "purpose" of the use, it is also transformative. In an immediate sense, the engineering design and purpose is to provide the LLM with the capacity to "generalize" from the training set as a whole. SUMF ¶ 9. That is what allows the LLM to predict the next token in any given sequence of text, including sequences it has never seen before. SUMF ¶ 9; Lafferty ¶¶ 70-71, 73-74, 152, 154, 161. The undisputed evidence shows that training's purpose is to draw out, then sharpen, the relationships that reflect the structure of human language and knowledge, including semantics, grammar, facts, ideas, associations, and so forth. SUMF ¶¶ 4.d-e, 13. And there is no dispute that every

step in the training process—from data acquisition, to dataset creation, to training—is directed to that transformative end. SUMF ¶¶ 9-12; Lafferty ¶¶ 89-90; *see Kadrey*, 788 F. Supp. 3d at 1048; *see Authors Guild*, 804 F.3d at 217 (recognizing that both the "making of the digital copies" for building a search engine *and* "use of those copies to offer the search tool were fair uses") (citing *HathiTrust*, 755 F.3d at 105).

Ultimately, the purpose of training an LLM is to create an innovative new technology for communication, speech, and the spread of knowledge. *See Oracle*, 593 U.S. at 30-32 (copying transformative because it served to develop "a highly creative and innovative [new] tool"); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992) (considering "ultimate purpose" of creating compatible video games). LLM technology can field virtually infinite, previously unseen queries and respond with novel analysis and expression. SUMF ¶¶ 9, 13; *see* Hurst Ex. 81, 274:24-276:13. An LLM can be built into existing productivity tools to make them better—indeed, among Microsoft's aims in partnering with OpenAI was to gain access to technology to enhance its own software and hardware. Hurst Ex. 126, 28:11-20, 30:2-8; Hurst Ex. 81, 170:9-14. And LLMs will yield brand new tools as well, promising a new generation of progress. For example, LLM technology is already helping provide students in developing nations with round-the-clock personalized tutors, speeding the work of under-resourced non-profit organizations, and assisting scientists with cutting-edge research. Hurst Ex. 121, 96:2-97:17, 99:25-101:15; Hurst Ex. 49, 354:2-355:2.

b. Case law confirms that training is transformative. The use of copyrighted works to train an LLM fits easily among transformative uses courts have found to be fair use as a matter of law.

For example, in *Authors Guild*, the Second Circuit held that Google Books is fair use. Google quite literally made a digital database of exact copies of "20 million books," "retain[ing] the original scanned image" of every page, to enable users to (a) search for books and (b) view word-for-word snippets from them. 804 F.3d at 208-10. Despite the database of exact copies and deliberate verbatim display of portions of works to consumers, *Authors Guild* found both that the search function was "highly transformative" and that the snippet view added "important value" to that use. *Id.* at 217-18. If the use of books to make a book-search tool is fair use as a matter of law, surely so too is the use of books to train an LLM—a speech and knowledge tool with a far broader range of uses. *Supra* 12-13.

In *Google v. Oracle*, the Supreme Court considered whether Google's copying of Oracle's Java "declaring code" into its Android-phone operating system was fair use. 593 U.S. at 28-30. The character of Google's use was not transformative at all. Google copied "11,500 lines of code" verbatim into its own operating system, *id.* at 33, and "it did so in part for the same reason that [the plaintiff] created those portions, namely, to enable [computer] programmers to call up implementing programs that would accomplish particular tasks," *id.* at 30. But the purpose of Google's use carried the day. The copying "expand[ed] the use and usefulness of Android-based smartphones," "offer[ed] programmers a highly creative and innovative tool for a smartphone environment," and ultimately facilitated "new products." *Id.* The purposes of LLM training are far more transformative, as is the character of the use. *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 923 (2d Cir. 1994) (contrasting "untransformed duplication[s]" with uses that "mak[e] some contribution of new intellectual value"). Again, if the copying in *Oracle* is transformative, LLM training is, too.

Also illuminating are two Ninth Circuit cases finding fair use as to "intermediate copying" of works for purposes of developing a technology. *Sega*, 977 F.2d at 1517, 1527; *Connectix*, 203 F.3d at 608. The defendants copied protected software "for the purpose of gaining access to the unprotected elements," enabling them to reverse engineer code that would allow for creation of new (non-infringing) works. *Connectix*, 203 F.3d at 602; *Sega*, 977 F.2d at 1520. In *Connectix*, for example, a software company copied the operating system of the Sony PlayStation "in order to find out how … PlayStation worked" and to develop a software program capable of "emulat[ing]" its non-expressive functional elements. 203 F.3d at 598-99. The Ninth Circuit deemed the technical character of the use transformative, even while finding the end product only "modestly" so. *Id*. at 606.

LLM training is likewise directed not towards use of any individual work's creative expression, but towards extracting underlying, unprotected elements of language and knowledge reflected across the whole training set. SUMF ¶¶ 11-13. It is fundamental that ideas, concepts, principles, semantic structures, writing styles, and facts are not copyrightable. § 102(b); *McDonald v. West*, 138 F. Supp. 3d 448, 459 (S.D.N.Y. 2015), *aff'd*, 669 F. App'x 59 (2d Cir. 2016). In effect, OpenAI used Books Plaintiffs' works—along with the rest of its massive dataset—to reverse engineer the functional components that enable the act of expression itself. And unlike in *Sega* or *Connectix*, where the resulting technology was a "modestly transformative" vehicle for making new video games, *Connectix*, 203 F.3d at 599, 606, LLMs may well be "the most transformative [technology] many of us will see in our lifetimes," *Bartz*, 787 F. Supp. 3d at 1033.

c. The commercial nature of training does not weigh against fair use.

This transformativeness is not diminished by the fact that OpenAI's purpose in training LLMs and Microsoft's purpose in partnering in model development is "of a commercial nature." § 107(1). "[M]any, if not most, secondary users seek at least some measure of commercial gain from their use[.]" *Texaco*, 60 F.3d at 921. All of the defendants in the fair use cases canvassed above, *supra* 14-15, had a commercial purpose. To prevent consideration of commercial purpose from "virtually obliterat[ing]" the fair use doctrine, *Texaco*, 60 F.3d at 921 (citation omitted), the Second Circuit has "repeatedly rejected the contention that commercial motivation should outweigh a convincing transformative purpose and absence of significant substitutive competition," *Authors Guild*, 804 F.3d at 219; *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 478 (2d Cir. 2004) (commercial nature "properly discounted" where use is "substantially transformative").

Here, Defendants' commercial purpose is outweighed by the transformativeness of the use. Any commercial benefit from LLM training is the result of a highly transformative technology that is itself the result of an extraordinarily expensive and resource-intensive project. Lafferty ¶ 44. And none of it comes from use of Plaintiffs' "precise form of expression [to] advance [Defendants'] commercial interests"—such as "using well-known copyrighted lines to attract attention to an advertisement." *Consumers Union of U.S., Inc. v. Gen. Signal Corp.*, 724 F.2d 1044, 1049 (2d Cir. 1983). The value is in trillions of tokens' relationships to one another. The "attenuat[ion]" of commercial value from the expression in any work makes Defendants' commercial purpose largely irrelevant to the first factor analysis. *Blanch v. Koons*, 467 F.3d 244, 262 (2d Cir. 2006) (Katzmann, J., concurring) (quoting *Texaco*, 60 F.3d at 922).

2. Plaintiffs' arguments based on "shadow libraries," "memorization," and the Bing Index do not weigh against the transformativeness of LLM training.

Facing the reality that every aspect of LLM training is transformative, Plaintiffs avoid confronting the training process as a whole, instead slivering off aspects of that process to posit narrow liability theories that might slip past summary judgment. First, they point to OpenAI's downloading of copyrighted books from so-called "shadow libraries"—locations on the internet where anonymous third parties have placed pirated content. Second, they tout an incidental, unintended byproduct of the training process—that a finished LLM can, through elaborate prompting, be made to generate text that overlaps with text in the training set. And third, they point to Microsoft's provision to OpenAI of portions of the Bing Index—the web index that powers Microsoft's Bing search engine. None of this materially alters the fair use analysis or creates an issue for trial.

a. Downloading from "shadow libraries." Like other challengers of generative AI, Books Plaintiffs focus on certain sources OpenAI used to obtain copyrighted books for training some of its earlier models. Most prominent is a website called Library Genesis, or LibGen, an online repository known to contain unauthorized copies of books. Hurst Ex. 22, 19:20-20:6. Plaintiffs seize on dicta from *Bartz* positing that "downloading source copies from pirate sites" like LibGen might be "inherently, irredeemably infringing." 787 F. Supp. 3d at 1025-26. Though *Bartz* did not ultimately adopt this categorical principle, Plaintiffs use it to theorize a separate "downloading claim." Hurst Ex. 131 at 9; Hurst Ex. 135; Ex. 146 at 5.

To begin with, this theory could never support a claim for *direct* infringement against Microsoft. A direct infringer is one who engages in the "*volitional conduct* that

cause[d] the copy to be made." *Cartoon Network LP, LLLP v. CSC Holdings, Inc.* ("*Cablevision*"), 536 F.3d 121, 131 (2d Cir. 2008) (emphasis added); *Fox Broad. Co. v. Dish Network L.L.C.*, 747 F.3d 1060, 1067 (9th Cir. 2014). In simple terms, direct infringement claims lie only against an actor who "actually 'makes' a copy"—"actually commits a copyright-infringing act." *Cablevision*, 536 F.3d at 131, 133. The undisputed factual record establishes that Microsoft did not download—and indeed had no involvement at all in acquiring—the datasets from LibGen that OpenAI used in connection with the development of its early models. SUMF ¶¶ 30-31. For that reason alone, this Court should grant summary judgment against any direct infringement claim against Microsoft predicated on OpenAI's downloading from LibGen.

But Books Plaintiffs' separate downloading claim also fails for a more fundamental reason: The purpose of downloading works from LibGen is the same as the purpose of *every other* use in the training process—from acquisition, to training, to the finished model. As *Kadrey* explained, downloading from a shadow library "must still be considered in light of its ultimate, highly transformative purpose: training [LLMs]." 788 F. Supp. 3d at 1047; *see Authors Guild*, 804 F.3d at 208 (analyzing together discrete acts of "scan[ning], render[ing] machine-readable, … index[ing] …, stor[ing]," and ultimately "retain[ing]" books for purposes of "enabl[ing] the Google Books search engine"). As long as the use at issue is "reasonably necessary to achieve," *Warhol*, 598 U.S. at 532, and "tethered to," *Oracle*, 593 U.S. at 34, the transformative purpose of training an LLM, the analysis is materially the same. And here, there is no question that downloads from LibGen carry that common objective. *See* SUMF ¶ 4.a; Hurst Ex. 111, 114:23-115:2.

To see that Plaintiffs' theory is dubious, one need look no further than the seminal cases discussed above—none of which treat the source or provenance of the work as legally significant. Each describes *where* the defendants found the copies they ultimately used in a transformative manner as happenstance background, and nothing in the fair use analysis turned on it. This makes sense, because fair use is focused on the incentives attending *the use itself*, and not surrounding circumstances that do not change the purpose, character, amount, substantiality, or market effect of the use. Here, LibGen happens to be one source of data OpenAI used to build a dataset large enough to enable LLM training. Had OpenAI obtained the same books from some other source, that would change nothing about how they were used to train an LLM and what effect LLM training had.

Tossing about words like "piracy," "bad faith," or "illicit" does Plaintiffs no good either. It is doubtful that these concepts are even relevant to the fair use analysis. Decades ago, in *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539 (1985), the Supreme Court considered the "propriety of the defendant's conduct" as part of its fair use assessment, noting that the defendant had "knowingly exploited a purloined manuscript." *Id.* at 562-63 (quotations omitted). More recently, however, the Supreme Court has expressed "skepticism about whether bad faith has any role in a fair use analysis." *Oracle*, 593 U.S. at 32-33 (citing *Campbell*, 510 U.S. at 585 n.18). And in *NXIVM*, a split Second Circuit panel treated "bad faith" as a "subfactor," but found it outweighed by the transformativeness of the use. 364 F.3d at 478; *contra id.* at 485 (Jacobs, J., concurring). As the concurrence in *NXIVM* reasoned, the right to fair use is an economic concept based on incentives for creation and innovation, not a privilege "earned by good works and clean morals." 364 F.3d at 485 (Jacobs, J., concurring); *see*

Leval, *supra*, 103 Harv L. Rev. at 1110 ("Copyright is not a privilege reserved for the well-behaved."). After all, *any* use of a work as to which a defendant asserts fair use is, by definition, unauthorized and objected to by the copyright owner.

This does not mean Microsoft condones or approves of shadow libraries or online piracy. To the contrary, Microsoft works hard to combat piracy every day and has developed extensive policies to address it. Hurst Ex. 81, 170:24-173:23; Hurst Ex. 81.A. Existing law affords rightsholders every ability to take action against those who host these shadow libraries and put a stop to *their* infringing uses. *See, e.g.*, *Cengage Learning, Inc. et al. v. Doe et al.*, No. 23-cv-8136-CRM, ECF No. 36 (S.D.N.Y. Sept. 24, 2024) (granting injunction against Library Genesis). But when it comes to the uses at issue *here*, the record establishes that OpenAI acquired all training data for the transformative purpose of training an LLM. Because that data is used in materially the same manner and for materially the same purpose as data from any other source, the provenance of the data simply does not "move the needle" on the fair use analysis. *Kadrey*, 788 F. Supp. 3d at 1046-47.

b. "Memorization." Books Plaintiffs have also focused on the possibility that training data "(at least in part) can be accessed, recalled, and reproduced by the LLM" in response to a user's prompt. Compl. ¶ 65. This is sometimes referred to as "memorization," which is something of a misnomer because—as the allegation just quoted demonstrates—the concept is based on observing *generated outputs* of an LLM that resemble training data (i.e., regurgitation), rather than observing model properties themselves. But Books and News Plaintiffs seize on the label to argue that a finished LLM actually contains copies of works in the training set rather than billions of mathematical parameters, as indisputable evidence shows. *E.g.*, NYT Compl. ¶¶ 106,

177-78 (suggesting LLMs are derivative works of every work in the training set); NYDN Compl. ¶ 96 (same); CIR Compl. ¶ 87 (same).

Nomenclature aside, even if one were to take "memoriz[ation]" of training works for "granted," it would not make a difference to the analysis because use of books to train an LLM remains "spectacularly" transformative. *Bartz*, 787 F. Supp. 3d at 1021. First, it is undisputed that memorization and regurgitation of copyrighted works is not the purpose of training. As Microsoft and OpenAI's experts explain without dispute, regurgitation is not an intended feature. SUMF ¶ 14. The purpose of LLM training, rather, is for the LLM to develop the ability to *generalize* from materials in the training set. SUMF ¶ 9; *see Texaco*, 60 F.3d at 924 (focusing on the "dominant" purpose of a copying use to assess transformativeness). An LLM's value is in abstracting *away* from the training data. *See* Hurst Ex. 81, 82:1-8 (LLMs are "fundamentally … only valuable" if they can abstract away from training data); Lafferty ¶¶ 100, 273. That is why engineers in the training process actively work to *avoid* memorization. SUMF ¶¶ 4.e, 13; Lafferty ¶¶ 71, 74-76.

Second, some level of regurgitation in no way distinguishes training from other uses courts have found fair. As discussed above, *supra* 14, the Google Books product at issue in *Authors Guild* required ongoing use of a database of unauthorized copies of books. Likewise, in *Oracle*, Google deployed a word-for-word copy of Oracle's declaring code in its own Android operating system to ensure that programmers could use its platform just as easily as Oracle's platform. *Supra* 14. For web-based search engines to function, full-text copies of the entire internet must be made on a continuous basis and stored in an index for quick retrieval in response to user queries. 4 *Nimmer on Copyright* § 13F.14[B] ("wholesale duplication" of websites to create a search index is

fair use). Such uses of copyrighted content are nevertheless transformative, and therefore, fair use as a matter of law. The same analysis should apply here: Even if Books Plaintiffs could demonstrate real-world memorization of training works (which they cannot), this would not detract from the transformative nature of the use overall.

In any event, the record shows that memorization is more a hypothetical technical curiosity than a practical reality. To generate 30-word "regurgitations," Books Plaintiffs' expert, Dr. Shawn Shan, relentlessly prompted GPT models with an adversarial extraction protocol. Shan used approximately 5.3 million attempts to feed the LLM verbatim book passages hundreds of words in length, fewer than 1% of which produced any 30-word matches. Hurst Ex. 107, 116:14-17, 145:18-146:17; Hurst Ex. 242 ¶¶ 17, 21, 29. Shan used these prompts because they were what "generate[d] the desired outputs"—that is, he already had the text, decided what passage to try to extract, then used surrounding text to bombard the LLM until it emitted what he was looking for. Hurst Ex. 107, 116:18-117:16.

This proves at most that a highly skilled and very motivated computer expert can use stylized prompting techniques to force an LLM into regurgitating small amounts of text in a lab setting. Indeed, if anything, Shan's all-out siege on the LLM only proves that regurgitation of training works is neither the purpose nor the effect of LLM training. That is why Shan found virtually no regurgitation outside of his lab. He reviewed 8.2 million conversations from Microsoft's LLM-based Copilot product and found *24* Copilot responses that included 30 matching words. SUMF ¶ 15.b. That is a .00029% rate of regurgitation. And for 202 out of 212 of Books Plaintiffs' asserted works, Shan found no regurgitation in the logs. SUMF ¶ 15.a. That hardly undermines the transformative purpose of LLM training.

c. Bing Index. Last is Books Plaintiffs' focus on Microsoft's provision, at OpenAI's request, of portions of its Bing search engine index. SUMF ¶ 16. The evidence is inconclusive as to whether portions of the index Microsoft shared with OpenAI contained any of Plaintiffs' works. At best, Plaintiffs' expert purports to have identified between 8 and 24 of Plaintiffs' works in portions of an earlier and later transferred Bing Index, respectively. Hurst Ex. 107, 302:2-307:7.

In any event, this transfer is subject to the same fair use analysis as any other challenged act, because transfer of works to OpenAI was indisputably a part of training LLMs. SUMF ¶ 17. Plaintiffs' own expert conceded that "if you define 'training' broadly as contributing to a final product model," provision of the Bing Index was a step in that process. Hurst Ex. 107, 318:11-319:9. For early portions of the index, OpenAI used data for "computational analysis"; that is, OpenAI experimented on this data for purposes of determining whether "web data can be useful for improving [LLMs]" as training data. *See* Hurst Ex. 37.A at -97041; ECF 546-2 at 35:19-22. This constitutes a "reasonable first step toward training an LLM," *Kadrey*, 788 F. Supp. 3d at 1048, because one cannot train LLMs without knowing which types of data are useful, Hurst Ex. 5, 130:1-13; Hurst Ex. 104, 17:7-18:7; SUMF ¶ 7.

Nor would it matter if the Bing Index contained works scraped from shadow libraries or other illicit sources. As a general web search index, the Bing Index is made by crawling every corner of the internet through automated bots, not by targeting any particular source or type of content. SUMF ¶ 32. *Cf. Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1155, 1157, 1164 n.8 (9th Cir. 2007) (recognizing that search engine indexes may incidentally crawl websites with pirated works). In short, the undisputed

evidence shows that the purpose and character of providing the Bing Index was training of an LLM, and therefore highly transformative for the same reasons.

B. The nature of the underlying works and the amount of them used carry little weight because of the transformativeness of the use (Factors two and three).

The second and third factors carry little weight in the analysis here. The second factor examines the "nature of the copyrighted work," § 107(2); it "has rarely played a significant role in the determination of a fair use dispute," *Authors Guild*, 804 F.3d at 220. The third considers the "amount and substantiality … used," § 107(3); but "[c]omplete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original," *Authors Guild*, 804 F.3d at 220-21. In other words, the first and fourth factors drive the analysis where (as here) "the creative work is being used for a transformative purpose." *HathiTrust*, 755 F.3d at 98 (citation modified).

Even if the second and third factors were legally material here, they favor Defendants. Training does use full copies of works that are at the core of copyright's subject matter, but it is not an "unchanged use"—it breaks the works into tokens, then analyzes the patterns between them, translating these into model weights, SUMF ¶ 4.b. Any copying, moreover, is "necessary to gain access to … unprotected functional elements" of the works, *Connectix*, 203 F.3d at 603—like the basics of language, facts, ideas, and concepts that belong to the public. SUMF ¶¶ 9-13; *supra* 15.

C. Training an LLM does not cause substantial market harm in any cognizable market (Factor four).

The fourth statutory fair use factor considers "the effect of the use upon the potential market for or value of the copyrighted work." § 107(4). Extensive discovery proves that the use of copyrighted works to train LLMs has no impact on any "traditional, reasonable, or likely to be developed markets" for books. *Texaco*, 60 F.3d at 930. This is unsurprising given how profoundly transformative LLM training is. Factor four "is concerned with secondary uses that … usurp a market that properly belongs to the copyright-holder." *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 110 (2d Cir. 1998). Harms caused by transformative uses are disregarded because transformative use "by definition" does not "substitute[] for the original work." *HathiTrust*, 755 F.3d at 99; *Authors Guild*, 804 F.3d at 223 (noting the "close linkage between the first and fourth factors"). Because undisputed record evidence establishes that the use of books to train an LLM does not usurp any market properly belonging to Books Plaintiffs, factor four weighs heavily in favor of fair use. SUMF ¶¶ 21-26.

1. Training an LLM does not substitute for books in their traditional markets.

The traditional market for books is undisputed. Plaintiffs sell copies of books to readers through their publishers. SUMF ¶ 21. In addition to direct book sales, some Plaintiffs generate revenue from their copyrighted works through audiobooks, foreign distribution rights, and occasionally film and television adaptations. SUMF ¶ 21.

Neither the use of Books Plaintiffs' works to train an LLM nor the finished model have any relationship to those markets. OpenAI is not in the market for reading material, and no "potential purchaser[]" of *Cleopatra* or *The Street Lawyer* for their ordinary purposes "[could] opt to acquire" an LLM "in preference to" the book. *Authors*

Guild, 804 F.3d at 223; *contrast Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 194 (2d Cir. 2024) (characterizing notion that "if users can download an identical copy of a Work for free, few will pay to buy or lease Publishers' editions" as "common-sense"). The GPT models operate in a completely "distinct market[]": the market for generative AI technology. *Oracle*, 593 U.S. at 37 (no competing substitute where "there are two markets at issue").

The book authors who brought this suit confirmed the absence of traditional market harm. Jonathan Franzen, for example, testified that [redacted] Hurst Ex. 35, 46:5-8. David Henry Hwang stated he "do[es not] believe" Defendants' LLMs are "hurting the market for [his] plays." Hurst Ex. 55, 153:21-154:12. Plaintiffs' admissions are proven by analysis of sales of their works, which show no drop that could be attributable to the advent of Defendants' LLMs. SUMF ¶¶ 22-24.

The undisputed absence of harm in traditional markets for books should end the inquiry. We briefly address below Plaintiffs' legally untenable attempts to conceive of alternative forms of market harm.

2. The Court should reject Plaintiffs' arguments regarding a new AI training market and theory of market dilution from non-infringing works.

Because Microsoft's evidence demonstrates lack of harm in Books Plaintiffs' traditional markets, it was incumbent on Plaintiffs "'to introduce empirical evidence countering [that] showing.'" *Kadrey*, 788 F. Supp. 3d at 1056 (quoting *Patry on Fair Use* § 6.13). Books Plaintiffs failed to do so. Based on their focus in discovery, Books (and News) Plaintiffs appear to posit two alternative types of market harm. Hurst Ex. 131 at 16-17; Hurst Ex. 135; Hurst Ex. 146 at 8-9. First, they claim they suffered a loss of

the opportunity to license use of their works as training data, causing them economic harm. Second, Books Plaintiffs posit a "market dilution" theory: that if, downstream from LLM training, LLM-based products are used to create new, non-infringing works, those works might compete with the works Defendants used in LLM training. Neither of these harms is legally cognizable, and both rest on unfounded speculation.

a. Market for training data. Books Plaintiffs principally claim harm in *the loss of payment for training itself*. SUMF ¶ 25. As both *Bartz* and *Kadrey* explain, this argument fails because the market for "licensing … works for the narrow purpose of training LLMs" is "not one the Copyright Act entitles Authors to exploit." *Bartz*, 787 F. Supp. 3d at 1032. Whether a training 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.

Binding Second Circuit precedent confirms this view. In *HathiTrust*, book authors offered "a 'lost sale' theory" under which "every copy employed" in building Google's searchable books database "represents a lost opportunity to license the book for search." 755 F.3d at 99. The court held that "[t]his theory of market harm does not work," because "economic 'harm' caused by transformative uses does not count." *Id.* at 99-100; *supra* 25.

Books Plaintiffs cannot circumvent this rule by reframing their "lost sale theory" around a supposedly emerging "potential market" for LLM training data. "[I]t is a given in every fair use case that plaintiff suffers a loss of a *potential* market if that potential is defined as the theoretical market for licensing the very use at bar." *Oracle*, 593 U.S. at 38 (quoting 4 *Nimmer on Copyright* § 13.05[A][4]). If a copier's unlicensed use were always viewed as usurping this potential transformative market, then factor four would

only ever favor the copyright holder. *Texaco*, 60 F.3d at 929 n.17. Consideration of "potential" markets is therefore limited to those that the plaintiff *would have* developed—but has left "untapped"—irrespective of the defendant's own copying. *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 91 (2d Cir. 2014).

Also legally irrelevant is evidence that model developers have licensed content for LLM development. "[C]opyright owners may not preempt exploitation of transformative markets … by *actually developing or licensing others* to develop those markets." *Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 145 n.11 (2d Cir. 1998) (emphasis added). And it is a fallacy to "infer[] market harm" from the fact that others—or even Defendants—have shown a "willingness to pay" for such use. *Whyte Monkee Prods., LLC v. Netflix, Inc.*, 174 F.4th 761, 805 (10th Cir. 2026) (citation omitted). After all, they may have done so simply to avoid the disruption, expense, and risk associated with lawsuits like this one.

In any event, such sporadic licensing does not establish the existence of a "reasonably efficient, reasonably priced, [and] convenient [to] access" licensing market for the entirety of the training data needed to train an LLM. *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1279 (11th Cir. 2014). A single author's work provides only an infinitesimal fraction of the tokens used to train LLMs. SUMF ¶¶ 10-11, 28.a. To obtain sufficient data for training, a model developer must obtain works by millions of authors—each of whose works are automatically copyright protected the moment they are "fixed in [a] tangible medium." § 102(a). These authors are dispersed, difficult to identify, and costly to negotiate or contract with on an individual basis. SUMF ¶ 28.e. And there remains no settled, workable means of valuing any given piece of training data. SUMF ¶ 28.c; Tucker ¶¶ 111-15.

Books Plaintiffs' experts suggest that the building blocks of an LLM training market—like buyers with demand and sellers with supply—are present such that an "emerging market" is "on its way to being established." Hurst Ex. 62, 206:2-13. They point to scattered examples of licensing deals that enable use of copyrighted work in LLM training. *See* Hurst Ex. 62, 210:10-211:19. Again, because any such market would be one Books Plaintiffs are not entitled to monopolize, this evidence cannot create a material issue of fact. But anyway, these isolated deals are nowhere near enough to suggest that a market will develop to provide the enormous volume and variety of data needed for LLM training. As a result, any regime that requires LLM developers to negotiate licenses for training data would be an enormous barrier to innovation. SUMF ¶ 28; Tucker ¶¶ 146-50.

As a counter to this argument, both sets of Plaintiffs have also sought to develop an argument that their works have greater value than other types of works when it comes to training an LLM. This claim is factually dubious, Lafferty ¶¶ 97, 99-100, 315-16, 322, but immaterial anyway because the argument cannot be squared with the logic of fair use. It is like saying an English professor's use of a stanza of *Thirteen Ways of Looking at a Blackbird* to teach perspective and scale is somehow less a fair use because that work is so valuable in illustrating those poetic concepts.

Even if it were true that Books Plaintiffs' content is "more valuable" for training generally, the training process places no value on the *specific expressive aspects* of works. SUMF ¶¶ 9-13. It pools all tokens together to *generalize up and away from* particular expression to ideas. SUMF ¶ 12. So what if Plaintiffs' works are more valuable in contributing to that transformative purpose? At most, this would simply mean that their works contain tokens that reflect good English, useful facts, and

insightful ideas—none of which Plaintiffs own. *Baker v. Selden*, 101 U.S. 99, 100-01 (1879); § 102(b). If anything, that a work is valuable in enabling the creation of transformative LLM technology weighs *in favor* of fair use, not against it. *Warhol*, 598 U.S. at 532 (copying fair use if "justified" as "reasonably necessary to achieve [copier's transformative] new purpose").

b. "Dilution" from non-infringing books. Last, Books Plaintiffs assert market harm based on the concern that if people use LLM-based products or tools trained on their works to create new *non-infringing* books, those books might wind up competing with their works. Compl. ¶ 150. They draw support from one court that seemed intrigued enough to sua sponte explore a so-called "market dilution" theory of harm arising from the possibility that downstream, non-infringing works may "indirectly substitute" for the originals by "compet[ing] with the[m]." *Kadrey*, 788 F. Supp. 3d at 1051-57. The court in *Kadrey*, however, did not weigh such harms in its analysis because it found no evidence of them occurring. *Id.*

The notion that a technology that allows new authors to create competing, non-infringing works would weigh *against* fair use has it precisely backwards. New non-infringing creative works are a copyright *good*—they are the whole point of fair use. Copyright protects only the author's "particularized expression." *Mattel, Inc. v. Goldberger Doll Mfg. Co.*, 365 F.3d 133, 135-36 (2d Cir. 2004) ("The protection that flows from … copyright is, of course, quite limited."). Copyright holders are not entitled to an "undeserving monopoly" over the "individual creative efforts" of others. *Superior Form Builders, Inc. v. Dan Chase Taxidermy Supply Co.*, 74 F.3d 488, 492 (4th Cir. 1996). Indeed, copyright "encourages others to build freely upon the ideas and

information conveyed by a work." *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 350 (1991). This is the crux of the Constitutional copyright bargain.

That is why courts have recognized, and explicitly tolerated, the reality that "transformative" (i.e., non-substitutive) uses could have some indirect effect on "the potential market for the original." *Castle Rock Ent.*, 150 F.3d at 145 (citation omitted); *see Warhol*, 598 U.S. at 528 (defining "substitution" as "use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work"). And they have flatly rejected "[a]n attempt to monopolize the market by making it impossible for others to compete." *Sega*, 977 F.2d at 1523-24. Instead, evidence that a transformative technology will yield new competitive works weighs *in favor* of fair use. *See id.* (transformative use "facilitat[ed] the entry of a new competitor"); *see Connectix*, 203 F.3d at 607 (transformative use is "legitimate competit[ion]" and not cognizable "economic loss").

Regardless, Books Plaintiffs' vision of AI-generated works swamping their normal markets is purely speculative. *See Kadrey*, 788 F. Supp. 3d at 1056 (declining to weigh factor four against AI developer where authors failed to present more than "speculation" about market dilution harms). The fact that some new books are being written with the help of LLM-based products does not establish that such books *will meaningfully compete* with the works of accomplished authors like Books Plaintiffs. Tucker ¶¶ 61-76. Quite the opposite, economic evidence shows that merely "increasing the number of available books does not mechanically reduce demand for existing works." Tucker ¶ 74. Research supports this view: The overwhelming majority of consumers would not purchase an AI-generated book—even if it were drastically less

expensive than an ordinary one. SUMF ¶ 26.c; Tucker ¶ 87. Little surprise, then, that none of Books Plaintiffs point to evidence of a lost sale.

Books Plaintiffs seemingly agree. Several have themselves conceded that they do not expect consumers will prefer AI-generated works as an alternative to Plaintiffs' own expressive works. *See* SUMF ¶ 26.d; Hurst Ex. 40, 47:21-48:6; Hurst Ex. 4, 250:25-251:8. And if generative AI does prove a useful tool in authoring new books, no one would be better positioned to harness this power than the incumbent authors who bring this lawsuit.

***

With strong opinions, entrenched economic interests, and choruses of likely amici on both sides, it is easy to see this case as the elder Byrons might have—pitting literary pursuits and scientific innovation against each other. But that is a false choice, as established law and the undisputed record in this case make plain. LLM training uses copyrighted works in transformative ways to advance the aims of copyright, without undermining any market that authors of creative works may properly exploit. Copyright interests are doubtless important, but they cannot be wielded to prevent technological progress, or to extract rents on innovation that only the largest companies could even fathom bearing. This Court should hold that LLM training is fair use as a matter of law.

To be clear, such a holding does not mean that *every downstream use* of an LLM—every LLM-based product, every end-user contrivance—is lawful. That the foundational technology underlying the internet is lawful does not mean every website and app is too. Likewise here, every LLM-based tool must be considered in light of its own design and function. As a case in point, Microsoft's brief in the News Cases explains why the specific LLM-based tool News Plaintiffs challenge is lawful. Tools

challenged in future cases may not be. But this Court should hold that the use of copyrighted works to train an LLM—the core analytical engine powering this technological revolution—is lawful. And it should reject attempts to muddy that clear conclusion with immaterial facts or doomsday rhetoric that have no place in a copyright case.

CONCLUSION

The Court should grant summary judgment in favor of Microsoft on all claims asserted against it.

Dated: September 4, 2026

Respectfully submitted,

/s/ Annette L. Hurst

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

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

mrshapiro@orrick.com
vsorrentino@orrick.com
hobeng@orrick.com

Sheryl Koval Garko (admitted pro hac vice)
Laura Brooks Najemy (admitted pro hac vice)
222 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 WORD COUNT COMPLIANCE

Pursuant to Local Civil Rule 7.1(c), the foregoing Memorandum of Law complies with the Court's order of August 7, 2026 (ECF 1664), as it contains 9,699 words, according to the word processing program used to prepare it.

/s/ Annette L. Hurst
Annette L. Hurst
Attorney for Defendant
Microsoft Corporation