EXHIBIT A

UNITED STATES DISTRICT COURT

SOUTHERN DISTRICT OF NEW YORK

IN RE OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This document relates to: *All Cases*	Case No. 25-md-3143-SHS-OTW Hon. Sidney H. Stein

PROPOSED BRIEF OF AMICUS CURIAE THE MINDEROO CENTRE FOR TECHNOLOGY AND DEMOCRACY, UNIVERSITY OF CAMBRIDGE IN OPPOSITION TO DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT

TABLE OF CONTENTS

Page

INTEREST OF AMICUS CURIAE ***1***

INTRODUCTION ***1***

SUMMARY OF ARGUMENT ***3***

LEGAL STANDARD ***4***

ARGUMENT ***5***

I. The Nature of Defendants' Use of the Asserted Works is a Disputed, Material Fact. **5**

A. The inquiry is objective and specific to the use. 5

B. Design goals are not the use. 6

II. The Nature of "Training" Over Time is a Disputed, Material Fact. 8

A. Training is cumulative, on Defendants' own account. 7

B. Books are presented again, and their contents are carried forward even when they are not. 7

C. That is why retention is material: an encoding that is carried forward is not an intermediate copy. 10

D. Microsoft's "even assuming memorization" fallback assumes a fact it cannot describe. 11

III. What the Models Retain and Use is a Disputed, Material Fact. **12**

A. Defendants offer output counts and generalities, and they concede the phenomenon. 12

B. The record contains substantial evidence that work-specific expression persists. 13

C. The output evidence comes from products built to suppress outputs. 14

IV. The Contribution of Each Asserted Work to the Accused Models is a Disputed, Material Fact. **15**

V. Microsoft's market evidence leaves a competing record unaddressed. **17**

CONCLUSION ***18***

TABLE OF AUTHORITIES

Page(s)

Cases

Allen v. Coughlin,
64 F.3d 77 (2d Cir. 1995) 5

Anderson v. Liberty Lobby, Inc.,
477 U.S. 242 (1986).....4

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith,
598 U.S. 508 (2023).....5

Am. Geophysical Union v. Texaco Inc.,
60 F.3d 913 (2d Cir. 1994).....7

Authors Guild v. Google, Inc.,
804 F.3d 202 (2d Cir. 2015).....11

Authors Guild, Inc. v. HathiTrust,
755 F.3d 87 (2d Cir. 2014).....7

Bartz v. Anthropic PBC,
No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025).....10, 11

Basic Books, Inc. v. Kinko's Graphics Corp.,
758 F. Supp. 1522 (S.D.N.Y. 1991) 7

Campbell v. Acuff-Rose Music, Inc.,
510 U.S. 569 (1994).....5

Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.,
150 F.3d 132 (2d Cir. 1998).....16

Google LLC v. Oracle Am., Inc.,
593 U.S. 1 (2021).....5

Hachette Book Grp., Inc. v. Internet Archive,
115 F.4th 163 (2d Cir. 2024) 17

Harper & Row, Publishers, Inc. v. Nation Enters.,
471 U.S. 539, 560 (1985) 5

Marvel Characters, Inc. v. Simon,
310 F.3d 280, 286 (2d Cir. 2002) 5

Morales v. Quintel Ent., Inc.,
249 F.3d 115 (2d Cir. 2001).....4

Nichols v. Universal Pictures Corp.,
45 F.2d 119 (2d Cir. 1930).....16

Salinger v. Colting,
607 F.3d 68 (2d Cir. 2010).....16

Sega Enters. Ltd. v. Accolade, Inc.,
977 F.2d 1510 (9th Cir. 1992).....10

Sony Computer Ent., Inc. v. Connectix Corp.,
203 F.3d 596 (9th Cir. 2000).....10

Tolan v. Cotton,
572 U.S. 650 (2014).....4

Statutes and Rules

17 U.S.C. § 101.....5, 15

17 U.S.C. § 107.....5

Fed. R. Civ. P. 56.....*passim*

INTEREST OF AMICUS CURIAE

The Minderoo Centre for Technology and Democracy at the University of Cambridge, affiliated scholars, Director Gina Neff, and Taryn J.M. Ward submit this proposed brief as amicus curiae. The Centre is the proposed Amicus; named scholars are listed in their personal capacities. Amicus examines how claims about what artificial intelligence systems contain and do are made, tested, and audited, and how those systems affect creative work.

That is the question on which Defendants' motions turn. Defendants ask the Court to hold that training their models on Plaintiffs' books was fair use as a matter of law. That holding depends on a factual account of what training does with a book. Amicus submits this brief to explain why that account is not undisputed on this record. Amicus has no financial interest in the outcome. No party or party's counsel authored this brief in whole or in part, and no person other than amicus contributed money intended to fund its preparation or submission.

INTRODUCTION

Microsoft and OpenAI's instant Motions for Summary Judgment (together, the "Motions") are about whether Defendants' copying the full text of Plaintiffs' books to train large language models was fair use. Defendants claim that training *is* fair use and so contend that summary judgment is appropriate insofar as no one disputes that the Asserted Works were used to train their models. But while there is no dispute that model "training" is how all parties define what Defendants did with the Asserted Works, there is a substantial and unresolved dispute about what "training" includes, and what happens to the Asserted Works during training. That dispute – the central dispute of the fair use argument – is material, factual, and the basis for denying the Motion.

The record evidence before the Court establishes that every Accused Model was built by converting the text of the Asserted Works into tokens and using those tokens to adjust a single set

of numerical parameters, millions of times over, with each adjustment computed from the state the earlier ones left behind (OAI SUMF ¶¶ 40–44; MSFT SUMF ¶ 4(d)).[1] Suppose *Anna Karenina* is among the first books a model trains on. By the end of the run, the model has trained on millions of other texts. Every one of those later steps was computed by a model that *Anna Karenina* had already helped to shape. Whatever the model took from Tolstoy's novel does not leave when the next book arrives; in some form, and to some extent, it is present in every step of training that follows, in the model that ships, and in every model later built from it.

"Training," then, is not a discrete act; it is one component of a longer chain of acts. What remains of each book, in what form, and what it does in every training run that follows are questions of fact. They are material, because Defendants' own theory of fair use depends on the answers. Defendants defend training as intermediate copying that extracts only unprotected, general features of language and leaves nothing of the authors' expression behind (MSFT Br. 15). But if the models keep the works' expression and carry it forward through every later step and into the products Defendants sell, that defense fails on its own terms.

On this record those questions are disputed. Defendants attempt to argue away the dispute with statements about design intent and accounts of how rarely their filtered consumer products emit long verbatim passages. But the record demonstrates that these questions are, as yet, unresolved. OpenAI's engineers wrote that its networks "can see a book a handful of times and they'll just memorize it" (SUF ¶ 560). OpenAI's own evaluations recorded its models regurgitating

[1] "OAI SUMF" refers to OpenAI's Rule 56.1 Statement (No. 1:25-cv-04315 Dkt. 598); "MSFT SUMF" to Microsoft's Rule 56.1 Statement (No. 23-cv-10211, Dkt. 780); "SUF" to Class Plaintiffs' Corrected Rule 56.1 Statement (Dkt. 1987); and "News SUF" to the News Plaintiffs' Corrected Combined Rule 56.1 Statement (Dkt. 2036). "MSFT Br." refers to Microsoft's Memorandum of Law (Dkt. 1690), "Pls. Br." to Class Plaintiffs' Memorandum of Law (Dkt. 1881), and "Lafferty Decl." to the Declaration of John D. Lafferty, Ph.D. (No. 24-cv-4872, Dkt. 666).

Plaintiffs' works (SUF ¶¶ 567–69). Defendants' own expert put it best: whether "the trained model behaves as though it retained the Asserted Works [is] the relevant empirical question" (Lafferty Decl. ¶ 282 (cleaned up)).

Amicus does not ask the Court to resolve that question and conclude whether and how the Accused Models retain and use Asserted Works during the training process. It asks the Court to recognize that these issues are unresolved, material, and fact-based. As such, the Defendants' Motions for Summary Judgment should be denied.

SUMMARY OF ARGUMENT

I. Fair use asks what the defendant actually did with the work, assessed objectively, use by use. Defendants' use is training, and it takes place inside the model. Defendants' design goals do not describe it, and neither do their users' purposes.

II. Training is a continuing process. Every Accused Model was built by cumulative adjustment of a single set of parameters, and books were presented to it repeatedly. Whatever a model takes from a book persists, to a disputed extent, through every later step, into the deployed model, and into later models built on it. That is why retention is material. A copy whose contents are carried forward into the finished product is not an intermediate copy, and Defendants' *Sega*/*Connectix* theory depends on there being nothing carried forward.

III. What the models retain of these works is disputed. Defendants offer output counts and general characterizations, and they concede that memorization occurs. The record contains substantial evidence that work-specific expression persists. The experts disagree on the central technical question. Defendants' own filings conflict with each other. And the output evidence Defendants rely on comes from products engineered to suppress the very outputs being counted.

IV. Books contributed either nothing specific to the work, in which case Microsoft's argument that books' value supports fair use falls away, or something specific to the work, which can reside only in the models' parameters. The record contains evidence for both. That is the dispute.

V. Microsoft's market evidence leaves a competing record about a different segment of the book market unaddressed. If dilution is cognizable, there is a live dispute.

LEGAL STANDARD

Summary judgment is appropriate only if "there is no genuine dispute as to any material fact." Fed. R. Civ. P. 56(a). The evidence is viewed in the light most favorable to the non-movant, with all reasonable inferences drawn in its favor. *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 255 (1986); *Tolan v. Cotton*, 572 U.S. 650, 657 (2014) (per curiam). On cross-motions, each motion is assessed on its own, and inferences run against each movant in turn. *Morales v. Quintel Ent., Inc.*, 249 F.3d 115, 121 (2d Cir. 2001). A genuine dispute of material fact exists when "the evidence is such that a reasonable jury could return a verdict for the nonmoving party." *Anderson*, 477 U.S. at 248. In making this determination, a court must "view the evidence in the light most favorable to the nonmoving party and draw all reasonable inferences in its favor, and may grant summary judgment only when no reasonable trier of fact could find in favor of the nonmoving party." *Allen v. Coughlin*, 64 F.3d 77, 79 (2d Cir. 1995) (internal citations and quotation marks omitted). "[I]f there is any evidence in the record that could reasonably support a jury's verdict for the non-moving party," summary judgment must be denied. *Marvel Characters, Inc. v. Simon*, 310 F.3d 280, 286 (2d Cir. 2002). inquiry of determining whether there is the need for a trial." Id. at 250.

Fair use is an affirmative defense, and Defendants bear the burden of proving it. *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994); *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 547 n.21 (2023).

ARGUMENT

I. The Nature of Defendants' Use of the Asserted Works is a Disputed, Material Fact.

A. The inquiry is objective and specific to the use.

Fair use is a mixed question of law and fact. Whether a use is fair is a legal conclusion, but what the use is, and what it does, are subsidiary questions of fact. *Oracle*, 593 U.S. at 24; *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 560 (1985). The Copyright Act directs courts to four factors in reaching that conclusion. 17 U.S.C. § 107.

The first factor asks about "the purpose and character of the use." 17 U.S.C. § 107(1). *Warhol* holds that the inquiry is objective and is directed to the specific use alleged to infringe. 598 U.S. at 532–33 & n.7, 544–45. The first step, then, is to identify what the use at issue actually is.

What is undisputed by both parties is that "training," generally, is the use in question. "To the extent that OpenAI downloaded any of Books Plaintiffs' Asserted Works, it was for training, not for other purposes" (OAI SUMF ¶ 104). But what does "training" mean, and what happens to the Asserted Works before, during, and after training? These are the fact questions that remain to be determined.

B. Use, Not Intent, Is the Central Question.

Defendants have defined training mainly by describing its aims. The "purpose of pretraining is to teach the model generalizable rules of language" (OAI SUMF ¶ 58). "[T]he fundamental objective of training is generalization" (MSFT SUMF ¶ 9). OpenAI "designs its

training algorithm to avoid memorization" (OAI SUMF ¶ 73). Microsoft's expert makes a similar point, positing that *development objective* in training is to improve the models' performance when presented with data it had not yet encountered, or "held-out" performance. (Lafferty Decl. ¶¶ 272–78; see OAI SUMF ¶¶ 45–48).

But describing what a model is designed *for* is not the same as defining what it *actually does*. Even if a warm stovetop and messy pans are not the design objective of cooking dinner, that does not mean they are not also part of the process. In the same way, the mere fact that the Accused Models may have been designed with the goal that they generalize language well does not provide a complete explanation of what else they do. Held-out performance measures whether a model generalizes. It does not measure whether the model also memorized its training text, and a model can do both. Dr. Lafferty's own sources say so. *See* Kushal Tirumala et al., *Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models* (2022) (cited at Lafferty Decl. ¶ 296).

The training signal that measures the process of training itself is fidelity to the author's text. That bears reiterating: the training signal, which both Defendants assert is a central value criterion of their training process, measures whether the Accused Model is able to predict the sequence of words from the Accused Work. At each step the model's prediction is compared with "the actual next word in the excerpt" (OAI SUMF ¶ 42), and the parameters are adjusted to make the author's actual word more likely (SUF ¶¶ 96–100). "This cycle repeats millions of times across many different excerpts, and the same model uses those excerpts to make a series of incremental adjustments to its parameters" (OAI SUMF ¶ 43; *see* MSFT SUMF ¶ 4(d)). Books are split into excerpts and taken in over many steps (OAI SUMF ¶ 37). OpenAI's engineers described the consequence plainly: "We train our networks to memorize the training data—that's their

objective" (SUF ¶ 560), and "models want to regurgitate, so we're inherently fighting against the training objective" (SUF ¶ 559). Whether that pull was overcome for these works, in these models, is a matter of fact, not design intent.

Notably, the Defendants support their argument about fair use by describing what their customers do with the Accused Models: tutoring, pro bono legal work, science, medicine (MSFT Br. 3, 13; OAI SUMF ¶ 22). A user may well make a transformative use of ChatGPT, but that says nothing about whether *Defendants'* own copying and training were fair. Outputs matter in this case only as evidence of what the models contain, and they are not the use under evaluation.

But while Defendants' first-factor case relies on what users do with the products, their fourth-factor case disclaims the same users. "[E]very LLM-based tool must be considered in light of its own design and function," downstream tools "may not be" lawful, and dilution through users' works is too remote and "has it precisely backwards" (MSFT Br. 30, 32–33). The same chain of downstream use is either part of Defendants' use or it is not, and that answer must be the same for both factors.

The Second Circuit has rejected borrowed purposes. "[A] use does not become transformative by making an invaluable contribution to the progress of science." *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 96 (2d Cir. 2014). Copying is not fair "simply because such copying is done by a company doing research." *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 924 (2d Cir. 1994). And a commercial copier could not rely on the educational purposes of the students who used its coursepacks. *Basic Books, Inc. v. Kinko's Graphics Corp.*, 758 F. Supp. 1522, 1531 (S.D.N.Y. 1991). Thus, when consumer usages of the Accused Models are properly set aside, what remains is the question of what and how the "training" use at issue here constituted.

Ascertaining whether *that* use -- the only use attributable to the Defendants here -- qualifies as fair use depends on the disputed facts in Parts II through IV.

II. The Nature of "Training" Over Time is a Disputed, Material Fact.

A. Training is cumulative, on Defendants' own account.

Defendants describe training as a process carried out on a single set of parameters. Those parameters begin as random numbers (OAI SUMF ¶ 40). As explained, the model takes an excerpt of text, predicts the next word, compares its prediction with the "actual next word," and updates its parameters (*id.* ¶¶ 41–42).

Each adjustment starts from the state the earlier ones left. Nothing in this process sets a book aside once its excerpts have been processed, and nothing in the record suggests that such discard takes place. Whatever an excerpt of an Asserted Work contributed to the parameters – which is, itself, a disputed, material fact -- is part of the model that processes the next excerpt, and the next book, and every text after that.

B. Books are presented again, and their contents are carried forward even when they are not.

The record shows two ways the use of a book continues after it is first processed.

The first use is re-presentation. OpenAI sampled its high-quality datasets more than once per training run and trained GPT-4 for multiple passes over some data (News SUF ¶¶ 871, 935–36). The order and amount of data are OpenAI's design choices (*id.* ¶ 933). Plaintiffs' statement records that "[c]opying repeats each time OpenAI passes the dataset through the model during a training epoch" (SUF ¶ 544).[2] OpenAI says it avoids repeating excerpts until a dataset is exhausted (OAI SUMF ¶ 74), but it concedes that repetition drives memorization (*id.* ¶¶ 67–68), and its engineers

[2]Each such pass may itself involve fresh reproductions of the work (SUF ¶¶ 535–44), but the Court need not reach that question to deny the motions.

wrote that models are "pretty good at memorizing docs — especially if we train for many epochs" (News SUF ¶ 954). How many times these books were presented to these models is itself disputed.

The second use is a kind of persistence. Even when a book is not in a batch of newly-ingested content, whatever the model has taken from the book sits in the same parameters that process the current text. Each later update either preserves it or overwrites it – the extent to which this takes place is an empirical claim, with research pointing in both directions. Deployed models, after all of their training, emit text from books they encountered during pretraining (Part III.B). Whatever such a model emits survived every step of training that came after the book was first processed.

The same is true across models. GPT-4 Turbo was "continued training of GPT-4" (News SUF ¶¶ 879–80). Microsoft's Copilot used a fine-tuned GPT-3.5 Turbo (*id.* ¶ 920). "Mid-training" continues training an already-trained model (*id.* ¶¶ 876–77). Whatever one model retained is the starting point for the next.

The picture that emerges is not "the book was read and the training for it is done." It is that each book, in some form, is part of every step of training that follows it, of the model that ships, and of the models built on that one. What form that is, and how much of the book it contains, is part of the decisive question in this case.

C. That is why retention is material: an encoding that is carried forward is not an intermediate copy.

Microsoft's theory is intermediate copying "to gain access to … unprotected functional elements" (MSFT Br. 15, 24, citing *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992), and *Sony Computer Ent., Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000)). Those cases rest on a particular sequence. A copy is made, it is used to reach unprotected elements, and then it is gone, leaving an end product that contains none of the copied expression. Defendants,

and the United States as an interested party, compare training to a person learning to write. The Department of Justice likens AI model training to Joan Didion typing out Hemingway's stories "to learn how the sentences worked" (Dkt. No 316 at 18). *Bartz* said that the models trained "[l]ike any reader aspiring to be a writer." *Bartz v. Anthropic PBC*, No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025) at *13.

The analogies also rest on an unstated premise: that a machine "learns" in a way close enough to a human writer to deserve the same treatment. Microsoft's own expert disclaims that premise. "[L]earning in the technical context is quite different from how humans learn" (Lafferty Decl. ¶ 16). OpenAI's statement puts "learns" in quotation marks (OAI SUMF ¶¶ 39), and Microsoft describes its model as "simulat[ing]" conversation and reasoning (MSFT SUMF ¶ 12). The Court need not decide whether machines learn as people do. If the analogy is offered as a factual claim, Defendants' own expert disputes it. If it is offered as a metaphor, it is not evidence. And the analogies hold only if training ends where Didion's typing ended. What Didion took from Hemingway remained in her mind, a domain that copyright law does not reach, unlike the commercial large language models owned by the Defendants.

More broadly, Defendants' reliance on *Sega*/*Connectix* would make sense if, after training on a given Asserted Work, the model ceased to make use of the Asserted Work ever again. But the record contains substantial evidence that whatever the model has taken from a work does not leave. It sits in the same parameters through which every later step, every later epoch, and every later model is computed (Parts II.B, III.B). *Sega*, *Connectix*, and *Google Books* each rested on a record that described the copy and what it did. This record does not. A copy whose contents travel into the finished product is not intermediate.

Whether that happened to these works is the disputed fact, and Defendants' own theory makes it decisive.

D. Microsoft's "even assuming memorization" fallback assumes a fact it cannot describe.

Microsoft argues in the alternative that even taking memorization "for granted," training remains transformative, because the copies in *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), were exact (MSFT Br. 21–22). But in *Google Books* the stored copy was described exactly. It was a full-text database serving a search index and limited snippets, and the court examined what a determined user could extract from it. *Id.* at 222–24. The court there could assess the precise use at issue because the record indisputably showed what was stored and what the stored copy did. Here, what the models hold of each work, in what form, and what it does in later training and in the deployed product are the facts that have not been established. An argument that assumes memorization and then asks the Court to evaluate a use nobody has described cannot carry a motion for summary judgment.

The fallback also runs into *Bartz v. Anthropic PBC*, No. 3:24-cv-05417-WHA (N.D. Cal. June 23, 2025), on which Defendants rely. Judge Alsup wrote that "[i]f this training process reasonably required making copies within the LLM or otherwise, those copies were engaged in a transformative use." Slip op. at 13–14. That observation was made on a record in which the authors did not allege infringing outputs and retention was not in dispute, and it rested on an analogy to human reading. It does not decide the question on this record, where retention is disputed and the evidence of it is substantial.

III. What the Models Retain and Use is a Disputed, Material Fact.

A. Defendants offer output counts and generalities, and they concede the phenomenon.

Every class-specific fact in OpenAI's Rule 56.1 statement about what its models contain is a count of outputs (OAI SUMF ¶¶ 82–91). So is Microsoft's (MSFT SUMF ¶ 15). The rest is generalization. The models encode "linguistic rules or common facts" (OAI SUMF ¶ 60), and their patterns are "properties of the aggregate" (MSFT SUMF ¶ 20). Microsoft's expert grounds his retention opinions in literature addressing "LLM behavior as a general phenomenon" (Lafferty Decl. ¶ 163). He has "every reason to believe" that OpenAI trained in the standard way (*id.* ¶ 278), and he offers no analysis of the Accused Models' parameters in relation to these works.

As set out above, Defendants concede that memorization occurs. "A model may encode common idioms or turns of phrase or even public-domain texts," and "[r]esearchers sometimes refer to this encoding of text sequences as memorization" (OAI SUMF ¶¶ 61–62). OpenAI's expert: "I did not and do not contest that memorization occurs" (Callison-Burch Reply ¶ 56, cited at News SUF ¶ 925). Microsoft concedes the models "can at times generate text that was part of their training corpus" (MSFT SUMF ¶ 14). Its expert calls memorization "a real but bounded phenomenon" that is more likely where text is repeated or contains "unusual or identifying" sequences (Lafferty Decl. ¶¶ 164, 293, 415). Published works are made of such sequences.

So the question is not whether memorization occurs. It is which works, how much, in what form, and with what role in the training that followed. On that question, Defendants' only work-specific evidence is outputs.

B. The record contains substantial evidence of replication and regurgitation

Defendants do not dispute that the Accused Models have, at times, reproduced text from Asserted Works (MSFT SUMF ¶ 14; OAI SUMF ¶¶ 83, 86). What they dispute is what that shows

about what the models hold. Files from OpenAI's Project Giraffe show its models regurgitating five Asserted Works, list eight more, and reflect internal searches that located ninety-six of Plaintiffs' works (SUF ¶¶ 567–69). One deployed model produced a contiguous 1,899-word excerpt of an Asserted Work (OAI SUMF ¶ 86). Ordinary production logs contain outputs reproducing 67 and 122 words (*id.* ¶ 83). One-sentence requests for a summary, with no book text in the prompt, returned verbatim passages of thirty words or more in six outputs (SUF ¶¶ 555–56). Defendants' logs show reproduction of seven Asserted Works across thirty-three responses and ten across twenty-four (SUF ¶¶ 550–51), and those logs largely omit the earlier periods before mitigations were deployed (SUF ¶¶ 552–53).

In fact, OpenAI's documents state that networks "can see a book a handful of times and they'll just memorize it" (SUF ¶ 560); that showing data once "causes some degree of memorization" (News SUF ¶ 943); and that "we should assume future big models will be able to memorize everything in the train corpus, even if it only appears once per epoch" (SUF ¶ 561). That last statement is a planning assumption, and it is presented here as one. OpenAI documents state that GPT-3.5 "can easily reproduce training data verbatim" (SUF ¶ 563; News SUF ¶ 959), that "current models … often regurgitate long sequences of copyrighted material" (News SUF ¶ 963), and that "the model will readily produce excerpts from works under copyright" (*id.* ¶ 966).

Class Plaintiffs' expert Dr. Zhao opines that training a language model is a form of compressing its training data (Zhao Rebuttal Report ¶¶ 14, 23–41, cited at Lafferty Decl. ¶ 424). The News Plaintiffs' expert Dr. Wenger opines that generalization requires memorization (Lafferty Decl. ¶ 416). Dr. Lafferty devotes a full section to disputing both (*id.* ¶¶ 415–31). Amicus takes no side in that debate here; its existence is the point. Qualified experts disagree, on this

record, about whether and how these models retain their training text. That is a genuine dispute of material fact, and it cannot be resolved on summary judgment by crediting one side.

C. The output evidence comes from products built to suppress outputs.

In response to this evidence of regurgitation, the Defendants point to guardrails and policies to mitigate the risk, but there is no answer at this stage about whether and how the Accused Models utilize the Asserted Works more generally after training. OpenAI's statement treats the whole subject as one of preventing regurgitation, not of what the models retain. It designs training "to avoid memorization that might lead to unwanted regurgitation" (OAI SUMF ¶ 73), refers to "problematic memorization" (*id.* ¶ 69), and post-trains its models "to avoid regurgitating any memorized training data" and to refuse requests (*id.* ¶¶ 76–77). OpenAI also deployed an inference-time filter to remove n-grams matching copyrighted material (SUF ¶ 557).

These are reasonable product choices, and amicus offers no commentary on them. That said, those choices bear on what the output evidence can show. A model can be trained not to emit memorized data only if there is memorized data to withhold, and a low output rate from a product designed to suppress output measures how well the suppression works, not what the model holds. The scale of the safeguards is itself evidence of what they guard against: post-training to avoid regurgitating "any memorized training data," trained refusals, and an inference-time filter keyed to specific copyrighted text (OAI SUMF ¶¶ 76–81; SUF ¶ 557). Safeguards on that scale would serve little purpose for a model that retained nothing. OpenAI's statement that ChatGPT "generally does not" quote books (OAI SUMF ¶ 105) describes trained refusal behavior (*id.* ¶ 77).

The timing matters too. ChatGPT launched without the filter (News SUF ¶ 1141; SUF ¶¶ 564–65). The first version covered roughly twenty books and *Harry Potter* (News SUF ¶¶ 1143–44; SUF ¶ 566). Copilot, for its part, grounds its answers in retrieved search results (MSFT

SUMF ¶ 15(c); Lafferty Decl. ¶¶ 18, 189–91), so its logs measure a retrieval system wrapped around a model, not the model.

Dr. Lafferty opines that training "does not store Plaintiffs' works as retrievable copies," and he denies that the parameters contain an index, a lookup function, or an "unzip" procedure (Lafferty Decl. ¶¶ 21, 279–83, 290–91). No one contends otherwise, and our argument does not turn on it. The Copyright Act asks whether a work can be "perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device." 17 U.S.C. § 101. It does not ask whether the work is indexed. Dr. Lafferty concedes that the parameters "may support instances of memorized continuations of text" (Lafferty Decl. ¶ 283), and that fine-tuning "can make memorized sequences easier to elicit" (*id.* ¶ 299). The question is what the models hold of these works and how it is used, and that remains disputed.

IV. The Contribution of Each Asserted Work to the Accused Models is a Disputed, Material Fact.

Microsoft argues that training "places no value on the specific expressive aspects of works" (MSFT Br. 29), and that books' value to training, such as it is, lies in "good English, useful facts, and insightful ideas," which supports fair use (MSFT Br. 29–30; MSFT SUMF ¶¶ 19–20).

Attribution is said to be impossible but somehow, simultaneously, also certified as undisputed. Microsoft states that it is not "currently technologically possible to isolate the contribution of any particular piece of data" (MSFT SUMF ¶ 28(a)). Its expert agrees: "it is impossible to ascribe responsibility for any particular model behavior to any one particular piece of training data" (Lafferty Decl. ¶ 335). Yet Microsoft asserts as undisputed that removing any work would not alter the model (MSFT SUMF ¶ 11), that no work confers capability (*id.* ¶ 10), and that the model's output is distinct from its sources (*id.* ¶ 13). Its expert concludes that the Asserted Works "had no meaningful impact on model performance" (Lafferty Decl. ¶ 322). If

attribution is impossible, those propositions cannot be established, let alone be undisputed. And the same expert who calls attribution impossible suggests that memorized text may come from third-party copies on the web (*id.* ¶¶ 289, 295), which is itself an attribution claim.

The record shows what books were sought for. OpenAI's first model paper valued books for their "long stretches of contiguous text, which allows the generative model to learn to condition on long-range information," and contrasted them with text "shuffled at a sentence level – destroying long-range structure" (SUF ¶ 450). OpenAI's corporate designee testified that books "contain very long, uninterrupted, correlated sequences of words" (SUF ¶ 453).

Shuffling a book's sentences preserves every word the author wrote -- what it destroys is the work's structure across its length. *Anna Karenina* opens: "Happy families are all alike; every unhappy family is unhappy in its own way." Any well-edited text can teach a model how one clause follows another. Only *Anna Karenina,* in its entirety, can contextualize that sentence and make it relevant across the eight hundred pages that follow it. Structure of that kind, when sufficiently particular, is protected expression. See *Salinger v. Colting*, 607 F.3d 68 (2d Cir. 2010); *Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132 (2d Cir. 1998); *Nichols v. Universal Pictures Corp.*, 45 F.2d 119, 121 (2d Cir. 1930). So are fictional facts and characters, which is what Dr. Lafferty means by the "names" and "associations" he acknowledges the parameters encode (Lafferty Decl. ¶ 297). *Castle Rock*, 150 F.3d at 139.

That produces a fork.

If books contributed nothing specific to the work, as Defendants' experts testify ("entirely fungible," SUF ¶ 497; MSFT SUMF ¶¶ 10–11; Lafferty Decl. ¶ 322), then nothing explains why Defendants sought books out and treated them as especially valuable (SUF ¶¶ 450, 453, 466–68, 480–83), and Microsoft's argument that books' value supports fair use has nothing behind it.

If books contributed something specific to the work, that contribution can reside only in the models' parameters. The models are fixed once deployed, and the base models retrieve nothing. Dr. Lafferty himself describes each book as its own statistical "regime," unlike "the full distribution of all human language" (Lafferty Decl. ¶ 428). And because books are split into excerpts for training (OAI SUMF ¶ 37), any knowledge a model has of a work as a whole had to be built up in the parameters across many excerpts and many steps. That is the continuing process of Part II, operating on a single book.

The record contains evidence for both prongs. The determination of which of those fact patterns describes these seven models and these works is a material fact that cannot be resolved on summary judgment. Amicus does not suggest that whole books were necessary to training because the operative question is what the copying captured and kept, not what training required.

V. Microsoft's market evidence leaves a competing record unaddressed.

Microsoft asserts that there is "no evidence of any such dilution of the market for well-known authors' books" (MSFT SUMF ¶ 26; MSFT Br. 9). In this Circuit the burden on the fourth factor stays with the party claiming fair use. *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 189 (2d Cir. 2024). The record contains competing evidence about a different segment of the book market. That evidence includes a study of 14,419 self-published genre-fiction titles finding that books with substantial AI-generated text rose from near zero to 31 percent of new top-25 entrants (SUF ¶¶ 858–67), and an OpenAI employee's prediction that "[t]he better we do on GPT-X, the more worried genre fiction authors will become about us substituting for them on Amazon" (SUF ¶ 749). Novelists themselves report their income declining as evidence of the

impact of AI on their ability to sell their original novels.[3] Amicus does not take a position on whether dilution is cognizable, but if it is, ¶ 26 is disputed.

CONCLUSION

Defendants ask the Court to decide that training was fair use on an account of training as a single, completed event. The record does not establish that account. Each book Defendants trained on became, in some form, part of every step of training that followed, of the models they deployed, and of the models built on those. What the models retain of each work, in what form, and how it is carried forward are material facts. Defendants have not established them, and the record puts them in dispute. The Court should deny Defendants' motions to the extent they seek a ruling that training on Plaintiffs' works is fair use as a matter of law.

Dated: October 30, 2026

Respectfully submitted,

MUCH SHELIST P.C.

/s/ James J. Ward

James J. Ward
MUCH SHELIST, P.C.
545 N.W. 26th St.
Suite 640
Miami, FL 33127
Tel: 786.581.6480

Attorneys for The Minderoo Centre for Technology and Democracy

[3]Collett, *The Impact of Generative AI on the Novel* (MCTD/IFOW, Nov. 2025) (survey of 258 published UK novelists).

CERTIFICATE OF WORD COUNT

In accordance with Local Civil Rule 7.1(c) and the Court's Order at Dkt. 1664, 25-md-3143, I certify that the foregoing brief contains 5,677 words, including footnotes but exclusive of the caption page, table of contents, table of authorities, certificates and signature block. The basis of my knowledge is the word count feature of the word-processing system used to prepare this brief.

/s/ James J. Ward