Exhibit A

UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF NEW YORK

IN RE OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION

This Document Relates to:
All Cases

Case No.: 1:25-md-03143 (SHS) (OTW)

Hon. Sidney H. Stein

[PROPOSED] ORDER BRIEF OF AMICUS CURIAE PROFESSOR MATTHEW SAG

Dated: October 29, 2026

Ethan Jacobs
Ethan Jacobs Law Corporation
100 Pine Street, Suite 1250
San Francisco, CA 94111
415.275.0845
ethan@ejacobslaw.com

Counsel for Amicus Curiae Professor Matthew Sag

TABLE OF CONTENTS

STATEMENT OF INTEREST OF AMICUS CURIAE 1

SUMMARY OF ARGUMENT 1

ARGUMENT 2

I. THE DEFENDANTS' PURPOSE AND CHARACTER WAS HIGHLY TRANSFORMATIVE AND STRONGLY FAVORS FAIR USE. 2

A. The evidence shows that defendants' AI training was profoundly transformative. 2

B. The transformative nature of LLM training is not undermined by vague assertions that LLMs "approximate" or "absorb" expressive value. 5

C. Transformativeness is measured by outputs under realistic conditions. 6

D. OpenAI's acquisition of training material from unsavory sources is a fourth factor consideration, not a distinct use. 9

E. The commercial nature of highly transformative uses is essentially irrelevant in these cases. 10

II. PLAINTIFFS ASSERT UNCOGNIZABLE AND LARGELY UNSUPPORTED HARMS. 10

A. There is no harm from expressive substitution. 10

B. There is no harm to copyright licensing markets. 11

1. Copyright owners cannot control markets for transformative uses. 11

2. Licenses granting access or avoiding litigation do not establish a broader market. 12

3. Content licensing at internet scale is impossible. 12

C. Individual rights are a poor fit for the speculative collective harms

described as "market dilution."....14

D. "Indirect expressive substitution" may be a cognizable market harm in appropriate circumstances....16

E. "Lawful access" may be relevant to the fourth factor, in appropriate circumstances....17

CONCLUSION....19

CERTIFICATE OF COMPLIANCE....20

TABLE OF AUTHORITIES

Cases

A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009) 2, 6
Am. Geophysical Union v. Texaco Inc., 60 F.3d 913 (2d Cir. 1994) 14, 19
Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023) 3, 4, 10, 12
Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) ..... passim
Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) ..... passim
Bartz v. Anthropic PBC, 787 F. Supp. 3d 1007 (N.D. Cal. 2025) ..... passim
Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605 (2d Cir. 2006) 12
Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) 3, 5, 11, 12
Castle Rock Entm't, Inc. v. Carol Publ'g Grp., Inc., 150 F.3d 132 (2d Cir. 1998) 12
Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) 16
Hoehling v. Universal City Studios, Inc., 618 F.2d 972 (2d Cir. 1980) 4
Kadrey v. Meta Platforms, Inc., 788 F. Supp. 3d 1026 (N.D. Cal. 2025) ..... passim
New York Times Co. v. Microsoft Corp., 777 F. Supp. 3d 283 (S.D.N.Y. 2025) 18
Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) 19
Romanova v. Amilus Inc., 138 F.4th 104 (2d Cir. 2025) 4, 12
Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992) 2, 4, 6
Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984) 4, 11
Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P., 756 F.3d 73 (2d Cir. 2014) 19
Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc., No. 25-2153, 2026 WL 2932557 (3d Cir. Sept. 29, 2026) 18

Statutes

17 U.S.C. § 102(a) 15
17 U.S.C. § 107 ..... 11, 16

Other Authorities

A. Feder Cooper et al., *Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models* 42 (July 20, 2026) (unpublished manuscript) 7
Ashley Gold, *Exclusive: FTC Urged to Investigate AI Firms for Destroying Books*, AXIOS (Aug. 21, 2026) 20
Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. 1163, 1209–10, 1214 (2025) 13
Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607 (2009) 5
Matthew Sag, *Copyright's Jagged Frontier*, 76 DUKE L.J. (forthcoming 2026) 10
Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1901 (2024)

..... passim

Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105 (1990) 3, 19

Zheng Dai & David K. Gifford, *Outputs of Generative Diffusion Models Are Often Unattributable*, 17 NATURE COMMC'NS, art. no. 6974 (Aug. 18, 2026)..... 16

Rules

Local Civil Rule 7.1(c) 20

STATEMENT OF INTEREST OF AMICUS CURIAE

I am the Jonas Robitscher Professor of Law in Artificial Intelligence, Machine Learning, and Data Science at Emory University School of Law. My research and teaching focus on copyright law and AI. For more than fifteen years my scholarship has addressed how the fair use doctrine applies to technologies that must copy works in order to analyze them. That scholarship has been cited by courts in the U.S. and overseas. I have no direct interest in the outcome of this litigation.[1] This brief supports neither party; my interest is in the coherence of the fair use doctrine as it is applied to a new technology.

SUMMARY OF ARGUMENT

Two district courts have now held, on summary judgment, that using copyrighted books to train a large language model is highly transformative, and ultimately, fair use. *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1021–22, 1033 (N.D. Cal. 2025); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1044, 1060 (N.D. Cal. 2025). Despite the controversy and seeming novelty of large language models (LLMs) and the astonishing scale of the copying in question, those decisions did not announce a new rule. They applied an understanding that has guided the copy-reliant technology cases in the Second Circuit and elsewhere for two decades: technical acts of copying that do not communicate an author's original expression to a new audience do not supersede the objects of the original and, in the absence of a significant countervailing market harm, are fair use. *Authors Guild v. Google, Inc.* (*Google Books*), 804 F.3d 202, 217–18 (2d Cir. 2015); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 639–40 (4th Cir. 2009); *Sega*

[1] No party's counsel authored this brief in whole or in part, and no party, party's counsel, or person other than amicus contributed money intended to fund its preparation or submission. Institutional affiliation is provided for identification only.

Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1522–23 (9th Cir. 1992).

On the record in this case, the defendants' use of plaintiffs' works for AI training is highly transformative and strongly favors a finding of fair use because under realistic conditions the outputs of defendants' models are in no relevant way similar to the invisible intermediate copies made during the training process. Against this, plaintiffs primarily assert harms under the fourth fair use factor that are uncognizable. There is no harm from direct expressive substitution in evidence, nor is there any cognizable harm to copyright licensing markets. However, to grant or deny the competing motions for summary judgement, the Court will need to assess narrower theories of harm based on (i) indirect expressive substitution that is likely to significantly undermine the incentives for creation of the specific works at issue or (ii) OpenAI's acquisition of training material from shadow libraries.

ARGUMENT

I. THE DEFENDANTS' PURPOSE AND CHARACTER WAS HIGHLY TRANSFORMATIVE AND STRONGLY FAVORS FAIR USE.

A. The evidence shows that defendants' AI training was profoundly transformative.

The first fair use factor, "the purpose and character of the use," asks in essence "whether the new work merely supersedes the objects of the original creation ..., or instead adds something new, with a further purpose or different character." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994) (cleaned up). Or, more succinctly, whether the use is transformative. *Id.*; *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 529 (2023). Transformative use plays a central role in our fair use jurisprudence because it both aligns the fair use doctrine with the constitutional imperative behind copyright, and helps courts make sense of the individual factors and their interconnection. Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990); *Google Books*, 804 F.3d at 214. For all that, courts have

sometimes struggled to understand how much transformation is enough—a particularly thorny question in appropriation art cases, but not in cases like those before the Court.

The Supreme Court's decision in *Warhol* clarified how the concept of transformative use should be applied by articulating what had previously been the doctrine's implicit rationale. *Warhol* explained that transformative use is "a matter of degree," 598 U.S. at 529, 532; that new expression "is not, without more, dispositive of the first factor," *id.* at 525; and that, ultimately, the benchmark of how much transformation is enough relates back to the threat of substitution: "the first factor relates to the problem of substitution—copyright's bête noire. The use of an original work to achieve a purpose that is the same as, or highly similar to, that of the original work is more likely to substitute for, or supplant, the work." *Id.* at 528 (cleaned up). In this context, the "substitution" the Court had in mind was substitution for original expression, as opposed to mere economic competition. A use that shares the purpose of a copyrighted work "is more likely to provide the public with a substantial substitute *for matter protected by the copyright owner's interests in the original work or derivatives of it*." *Id.* at 531–32 (cleaned up, emphasis added). The "*copyright owner's interests in the original work*" is another way of saying the copyright owner's original expression. *Id.*; *see also Google Books*, 804 F.3d at 214, 224 (focusing on "appropriation" that "serve[s] as a substitute for the original" and analyzing market harm in "the protected aspect of the author's work"). Thus, the substitution the Court was referring to in the context of the first fair use factor was expressive substitution, not competition in general. This makes sense: copyright encourages the production of competing works that do not trade on an original author's protected expression, but only copy facts and other information asserted therein. *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972, 974, 978–79 (2d Cir. 1980); *Sega*, 977 F.2d at 1523–24.

The Second Circuit has since explained how *Warhol*'s emphasis on "justification" fits this picture. In *Romanova v. Amilus Inc.*, 138 F.4th 104 (2d Cir. 2025), Judge Leval wrote that *Warhol* "made clear that *Campbell*'s requirement of justification … applies generally to all claims of fair use." *Id.* at 112. Transformative use is not the only path to justification under the first factor, *see Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 449–51 (1984) (private, noncommercial time-shifting); *HathiTrust*, 755 F.3d at 101–03 (access for the print-disabled), but it is the predominant one because the essence of truly transformative uses is that they are manifestly unlikely to substitute for the copyright owner's original expression—that is what makes them transformative.

Computational processes that employ copyrighted works as inputs at one end, and produce non-infringing outputs at the other end, are transformative—so long as they do so without communicating any meaningful portion of the original expression embodied in the inputs to any new audience. Indeed, they are not just transformative, but "spectacularly so." *Bartz*, 787 F. Supp. 3d at 1021; *see id.* at 1033 (the technology "was among the most transformative many of us will see in our lifetimes"); *Kadrey*, 788 F. Supp. 3d at 1044 ("no serious question" that training was "highly transformative"). Some refer to such processes as "nonexpressive uses,"[2] but nothing turns on that terminology.

Defendants[3] fed text into a computer system, where no human would ever read it, to build statistical models of language called LLMs. An LLM is not a library or a database. It is a set of

[2] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607 (2009) (coining the term "nonexpressive use" and arguing that nonexpressive uses are generally fair uses).

[3] For ease of exposition, this brief does not distinguish between actions taken by OpenAI, as opposed to Microsoft, except where relevant. This brief does not address the News Plaintiffs' separate claims against Microsoft concerning retrieval and "web grounding" in Copilot.

numerical parameters that capture patterns and structures emerging from the training data as a whole. In their ordinary operation the defendants' models respond to prompts with new text, not with copies of what they were trained on. Such a use does not "supersede the objects of the original creation" by substituting for its expressive content. *Campbell*, 510 U.S. at 579 (cleaned up). It derives uncopyrightable information from the collections of works and uses that information for new ends.

LLMs are new, but courts have faced similar nonexpressive uses many times and consistently found them to be transformative. Google's copying of millions of books served a "highly transformative purpose" because it told the public something about the books without supplying their expression. *Google Books*, 804 F.3d at 217–18; *see HathiTrust*, 755 F.3d at 97. Archiving student papers to detect plagiarism was "completely unrelated to expressive content." *iParadigms*, 562 F.3d at 639–40. Copying a program to build a competing, noninfringing alternative was "a legitimate, essentially non-exploitative purpose." *Sega*, 977 F.2d at 1522–23. These cases are not confined to computer code, nor do they hold that fair use is a rule of strict necessity. *See Google Books*, 804 F.3d at 221 (copying the totality of each book was "reasonably appropriate to Google's transformative purpose"); *HathiTrust*, 755 F.3d at 98 (copying the entirety of the works was "reasonably necessary" to enable full-text search and so not "excessive"). These rulings reflect both the intrinsic logic of copyright's idea-expression distinction and copyright's overall purpose in encouraging new works and new insights. Both favor defendants here.

B. The transformative nature of LLM training is not undermined by vague assertions that LLMs "approximate" or "absorb" expressive value.

The News Plaintiffs suggest that generative AI is somehow different from plagiarism detection, creating a book search index, or conducting statistical analyses on large corpora of

books, because training is a "consumptive use" that "exploits [the] intrinsic expressive value" of their works and "approximat[es]" their expression. ECF No. 1709 at 63; *see id.* at 59, 73–74. Learning valuable things from expression is not the same as reproducing that expression. Copyright "does not forbid the ordinary reader from extracting and reproducing the facts, ideas, or artistic techniques embodied in a work; it encourages them to do so." Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1901 (2024) [hereinafter *Fairness*]. What a language model derives from any text is which words (or more specifically, which tokens) tend to follow which, in what contexts, across billions of examples. Some of what LLMs learn is specific, such as the very high probability that Princeton was founded in 1746;[4] much of it is general, such as the fact that "hair" is a likelier next word than "propeller" after "the girl with the dark"; and much of it is truly abstract.

C. Transformativeness is measured by outputs under realistic conditions.

Although training and outputs are legally distinct acts, evidence about the general character of the latter is central to understanding the character of the former. Why? Because an AI model with a strong tendency to reproduce its inputs as outputs could hardly be described as transformative.

The extent to which LLMs memorize their training data is disputed, and I am not aware of any study which can credibly show how common memorization is for the average book or news article. Indeed, the *careful* studies I am familiar with expressly disclaim being able to make that calculation.[5] In broad terms, the research shows that some memorization occurs; it does not

[4] For an illustration, type "Princeton University was founded in the year" into https://matthewsag.com/nextword.html (a demonstration page that queries OpenAI's models through their programming interface and displays the probabilities the model assigns to candidate next words).

[5] *See, e.g.*, A. Feder Cooper et al., *Extracting Memorized Pieces of (Copyrighted) Books from*

show that memorization of the ordinary book or news article is common. Memorization is a phenomenon warranting investigation, not a general picture of how models treat their training data. The Court can decide summary judgment without making any specific finding in this regard because, although extensive memorization of training data inside an AI model *could* jeopardize the case that model training was transformative, whether it does so depends on what comes out of the model under real-world conditions. That is why the Second Circuit focused on security measures in *Google Books*, 804 F.3d at 227–28, and *HathiTrust*, 755 F.3d at 100–01, and why the courts in *Bartz* and *Kadrey* focused on the filters and "mitigations" that stood between the models and the public. *Bartz*, 787 F. Supp. 3d at 1018, 1021; *Kadrey*, 788 F. Supp. 3d at 1041–42.

In *Google Books*, the Authors Guild argued that a determined researcher could piece together a fraction of a book from snippets. 804 F.3d at 222–23. The Second Circuit was unmoved: snippet view, "at best and after a large commitment of manpower, produces discontinuous, tiny fragments." 804 F.3d at 224. The summary judgment record here is similar. Defendants' models do not, in their ordinary operation, deliver the plaintiffs' expression to anyone. The parties dispute how regurgitation in ChatGPT conversations should be counted, but on any measure it is rare. OpenAI's expert found 24 verbatim regurgitations of the News Plaintiffs' works in 20 million conversations, and the News Plaintiffs' own experts reported rates between 0.000002% and 0.00011%. ECF No. 1738 ¶¶ 219–24; ECF No. 1753 at 9. Even the News Plaintiffs' broadest measure, any 25-word overlap with any asserted work, reached some

---

Open-Weight Language Models 42 (July 20, 2026) (unpublished manuscript), https://arxiv.org/pdf/2505.12546v6 ("our experiments cover roughly 0.1% of the corpus. This is not nearly a large enough sample to make general claims about overall books extraction for the setup we study").

106,000 of the 20 million conversations, about one in two hundred. ECF No. 2195 at 198–99. A review of 8.2 million Copilot conversations found at least 24 containing a near-verbatim block of at least thirty consecutive words from the Books Plaintiffs' works, including matches in hidden web-grounding text never shown to users. ECF No. 1690 at 28; ECF No. 1974 ¶ 15; ECF No. 2226 at 22–24. The plaintiffs' own experts appear to have extracted only fragments, and only under heroic adversarial conditions that the public does not encounter and that would make no economic sense to replicate. Across roughly 5.3 million prompts, the longest contiguous passage Dr. Shan could extract from one of the most popular books of recent times was 1,899 words of *A Game of Thrones*, less than one percent of the novel. ECF No. 1738 ¶¶ 85–88; ECF No. 2195 at 104 (undisputed). The News Plaintiffs' own memorization studies submitted some 195 million and 65 million prompts and queries through OpenAI's developer interface, largely at temperature zero and largely to base models rather than the ChatGPT service the public uses. *Id.* ¶¶ 229–37; ECF No. 2195 at 211–15. By OpenAI's calculation, which the plaintiffs contest, the works they claim were regurgitated were reproduced on average to the extent of less than 4.4%. ECF No. 1738 ¶¶ 233, 238; ECF No. 2195 at 212, 216. No one spends millions of queries to recover a few pages of a novel that can be bought for the price of an airport sandwich, or a news article available on the publisher's own website.

The News Plaintiffs' experts describe memorization as parameters "encod[ing] copies of ... training data," ECF No. 1730 ¶ 925, but it is no answer to the lack of real-world outputs to describe LLMs as encoded or "compressed" copies. These descriptions are artful but misleading: they are true in a technical sense, but that sense has little to do with what the words "encoded" and "compressed" ordinarily convey. To most people, compression means MP3s, JPEGs, and ZIP files: formats that store every work by the same fixed rules and reconstruct it in a form

functionally indistinguishable from the original. AI training does neither. AI models learn whatever statistical patterns improve their predictions, and any true memorization that does occur is uneven, incidental, and unpredictable. *See* Matthew Sag, *Copyright's Jagged Frontier*, 76 DUKE L.J. (forthcoming 2026) (manuscript at 42–45) (Sept. 30, 2026 draft), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6319379.

D. OpenAI's acquisition of training material from unsavory sources is a fourth factor consideration, not a distinct use.

In *Bartz*, Judge Alsup found that Anthropic had assembled and retained a general-purpose research library of pirated books, keeping copies "should they prove useful for one thing or another," and held that building and maintaining that library "was its own use—and not a transformative one." 787 F. Supp. 3d at 1026. The holding turned on a standalone use with its own purpose that could not be justified as an integral part of AI training. In this case, without that extreme finding, splitting the ordinary steps of one integrated process into separate "uses" when they share a unitary purpose and character would be error. Judge Chhabria made the point in *Kadrey*: although Meta's downloading of books was "a different use from any copying done in the course of LLM training," it "must still be considered in light of its ultimate, highly transformative purpose: training Llama." 788 F. Supp. 3d at 1047–48; *see also id.* at 1060. *Warhol* did not analyze each step of Andy Warhol's production process; it analyzed the licensing use that was challenged. 598 U.S. at 533. The Second Circuit in *Google Books* described how Google "makes a digital scan of each book, extracts a machine-readable text, and creates an index," retaining the scanned images to improve the text later, 804 F.3d at 208, but it did not analyze the scan, the text extraction, the index, and the retained images as separate uses, each requiring its own fair use analysis. *See id.* at 216–18.

If the sourcing of material for AI training from shadow libraries moves the needle on fair

use, it should do so under the fourth factor. *See* Part II.E.

E. The commercial nature of highly transformative uses is essentially irrelevant in these cases.

Commercialism is not an equal and countervailing element to be weighed against the "purpose and character" of the use under factor one. As the text of Section 107 indicates, commercialism is merely one aspect of that use,[6] which varies in importance, and it is of no importance here. Courts have long recognized that "the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use." *Campbell*, 510 U.S. at 579; *see Warhol*, 598 U.S. at 531 (the commercial nature of a use "is not dispositive" and "is to be weighed against the degree to which the use has a further purpose or different character"). Nothing the Court said about Andy Warhol Foundation's non-transformative commercial use is to the contrary.

II. PLAINTIFFS ASSERT UNCOGNIZABLE AND LARGELY UNSUPPORTED HARMS.

A. There is no harm from expressive substitution.

The same thing that makes AI training transformative under the first factor in these cases probably defeats market harm under the fourth factor: there is no expressive substitution.

Expressive substitution is the line running through the first and fourth factors, but with a difference in emphasis. Some uses may offer potential for expressive substitution in theory, and thus lack transformativeness under the first factor, but if that potential is unlikely to be realized, the use may nonetheless be favored under the fourth factor, *Sony*, 464 U.S. at 449–51, especially if there is some other compelling non-transformative justification. *HathiTrust*, 755 F.3d at 101–

[6] Thus, the statute provides that the first factor is "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1).

03. Most of the time, however, the first and fourth factors run together because they both turn on, as *Romanova* put it, "interrelated issues contributing to a holistic inquiry." 138 F.4th at 117 n.9. Thus, the Court in *Campbell* noted that "when ... the second use is transformative, market substitution is at least less certain, and market harm may not be so readily inferred." 510 U.S. at 591. *Warhol* shows the same syllogism in reverse: the defendant's use conveyed significant aspects of the original expression captured in Goldsmith's portrait of Prince to a new audience, for substantially the same purpose as the original. 598 U.S. at 526.

Nonexpressive uses have such a strong track record in the fair use case law because a use that conveys none of the author's expression to a new audience seems inherently non-substitutive. There may be market effects in a broader economic sense, "but the 'market' and 'value' referred to in the fourth fair use factor are not simply any benefit the copyright owner might choose to nominate." *Fairness*, 92 Fordham L. Rev. at 1916.

B. There is no harm to copyright licensing markets.

1. Copyright owners cannot control markets for transformative uses.

The lack of harm from expressive substitution can't be offset by imagining a licensing market for such highly transformative uses. It is well-established that a copyright owner cannot control fair use markets merely "by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work." *Castle Rock Entm't, Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 146 n.11 (2d Cir. 1998); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (2d Cir. 2006). The Authors Guild argued that not being paid for text mining was a cognizable harm in *HathiTrust* and again in *Google Books*, and lost both times. *HathiTrust*, 755 F.3d at 99–100; *Google Books*, 804 F.3d at 223–27. *Bartz* and *Kadrey* followed suit. *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032.

2. Licenses granting access or avoiding litigation do not establish a broader market.

Many AI developers have entered into licenses for preferential access to content, and the record contains a catalogue of such agreements, most of whose terms are sealed. ECF No. 1730 ¶¶ 146–228, 274; *see also* ECF No. 1682 at 5 n.13. The fact that access to current and paywalled content, reliable delivery, and the right to retrieve and display material in response to user queries is worth paying for is encouraging, but it does not establish market harm from a general failure to license.

Such licenses show that channels of remuneration associated with generative AI already exist and will feed money back into the creative ecosystem. Declining to recognize that training on works scraped from the open internet is fair use, in the circumstances of these cases, would be more likely to destroy a market than to create one. As discussed below, no amount of judicial encouragement can conjure a licensing system for hundreds of millions of unidentifiable authors into existence. More likely, it would simply destroy an American industry and push the cutting edge of AI development to jurisdictions with more favorable copyright law (or places where private rights matter less than state imperatives). *See* Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 EMORY L.J. 1163, 1209–10, 1214 (2025).

3. Content licensing at internet scale is impossible.

Even if this Court determined that plaintiffs should be entitled to license the defendants' highly transformative use, it would have to contend with the impossibility of such a market forming. The problem in a nutshell is that although training data is valuable in aggregate, the marginal contribution of any individual piece of training data, whether a book or a social media post, is approximately zero. No amount of magical thinking will establish a market mechanism

under which defendants could transact with hundreds of millions or billions of rights holders. As such, rights clearance at internet scale is not merely difficult, challenging, expensive, it is literally impossible.

To elaborate, publicly disclosed training-set sizes for frontier models reached roughly 18 trillion tokens by late 2024. ECF No. 1690 at 12–13; ECF No. 1701 ¶ 5. The News Plaintiffs, who include some of the largest publishers in the country, assert, through their experts, that at least 6.3 million of their works were used in training. ECF No. 1738 ¶ 204. Assuming one thousand tokens an article, that is some six billion tokens: roughly three hundredths of one percent of such a corpus. Who wrote the other 99.96%? Hundreds of millions, or possibly billions of people and corporations, almost none of whom manage their copyrights, belong to a collecting society, or could be found at a cost that was not several orders of magnitude greater than their individual contribution. A market in which the price of each unit approaches zero and the cost of each transaction is more than nominal is not a market "traditional, reasonable, or likely to be developed." *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 930 (2d Cir. 1994).

Moreover, nothing in plaintiffs' theory of liability depends on any feature peculiar to their works. Plaintiffs claim that their copyrighted works were included, without permission, in a corpus used to train large language models. That is equally true of nearly every category of expressive work that has ever been published to the internet. Copyright attaches automatically upon fixation, without registration or notice, to any original work of authorship. 17 U.S.C. § 102(a). The training corpora at issue were assembled from broad crawls of the public web, which means they contain the work of a large fraction of the people who have written anything online in the past three decades. If the Court holds that unlicensed training is not fair use, that holding will apply with identical force to every other class of author whose work appears in the same

datasets. The Court should therefore expect follow-on class actions from, at a minimum:

1. authors of books outside the current classes;
2. authors of academic articles;
3. freelance authors of magazine and news articles;
4. authors of letters to the editor and of student newspaper articles;
5. authors of advertising copy;
6. authors of photographs and illustrations;
7. authors of blogs, subscription newsletters, and personal essays;
8. authors of serialized web fiction and fan fiction;
9. authors of wiki and fan-wiki entries;
10. authors of questions and answers on question-and-answer sites;
11. authors of forum and message-board posts, potentially organized platform by platform;
12. authors of social media posts; and
13. authors of product, restaurant, and travel reviews.

All these authors made an invaluable contribution to the training corpus *in the aggregate* because what makes training data valuable is scale and diversity, not any particular work or class of works.

C. Individual rights are a poor fit for the speculative collective harms described as "market dilution."

The "market dilution" theory of harm advanced by at least some of the plaintiffs in these cases should be rejected. The theory, as framed by Judge Chhabria in *Kadrey*, is that training AI models on plaintiffs' works would enable the downstream production of competing AI-produced and AI-assisted works. Such works, even if far removed from the plaintiffs' copyrighted material, could theoretically "indirectly substitute" for it, facing authors with a flood of competition. *Kadrey*, 788 F. Supp. 3d at 1051–57.

There are three obvious problems with the market dilution theory. First, non-infringing works that merely occupy the same genre or topic category as those of the plaintiffs are a strange place to look for harm. Copyright's idea-expression distinction encourages competition on that which is unprotected. The freedom to build on the ideas and information in a work is "not some unforeseen byproduct of a statutory scheme" but "the essence of copyright, and a constitutional

requirement." *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349 (1991) (cleaned up).

Second, there is no evidence of market dilution. The fourth factor asks about "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). To show dilution, a plaintiff would have to connect the defendant's use of her work to a change in the market for her work. Each link in that chain is missing and is likely to stay missing.[7] Because the contribution of any single work to a model trained on trillions of tokens is approximately zero, a model trained without the plaintiffs' works would behave in essentially the same way, and whatever its outputs do in the market cannot be attributed to the use of those works. Empirical research on generative image models illustrates the point: as training datasets grow, the influence of individual training images—or even collections of images by individual creators—on generated outputs diminishes. Zheng Dai & David K. Gifford, *Outputs of Generative Diffusion Models Are Often Unattributable*, 17 NATURE COMMC'NS, art. no. 6974 (Aug. 18, 2026), https://www.nature.com/articles/s41467-026-75667-5 (finding that, in diffusion models trained on sufficiently large datasets, removing particular sources often leaves the resulting outputs unchanged).

Third, speculation about collective harm is not evidence of harm tied to individual rights. In the non-speculative present, training LLMs on a diverse corpus of material scraped from the internet has produced an astonishing new technology, with benefits that will surely be addressed in other amicus briefs. Against this, the plaintiffs offer speculation as to future *collective* harms—and therein lies the problem. Copyright is a system of individual rights, but the harms asserted under the banner of dilution are collective harms. It is not clear what gives these

[7] Even if the Class Plaintiffs' principal evidence of dilution, a working paper that defendants showed was funded by Class Counsel, had not been struck from the record, *see* ECF No. 2046 at 1; ECF No. 2090, it would not have filled this gap.

individual plaintiffs a right to speak for the millions of authors who believe generative AI is largely beneficial or to monetize a veto right that, if it belongs to one author, belongs to all. This is precisely why such questions are best left to the representative branches.

D. "Indirect expressive substitution" may be a cognizable market harm in appropriate circumstances.

Although competition, in general terms, has not been treated as a cognizable harm under the fourth factor, indirect expressive substitution that is likely to significantly undermine the incentives for creation of the specific works at issue certainly could be. A court could reasonably find the systematic extraction of valuable uncopyrightable material from an information source (such as a website), used to create a direct market substitute for that information source, presented cognizable market harm under the fourth factor. Especially if such extraction was likely to significantly undermine the source's incentives for original content production. *Fairness*, 92 Fordham L. Rev. at 1919. The Third Circuit's recent decision in *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, No. 25-2153, 2026 WL 2932557 (3d Cir. Sept. 29, 2026), is consistent with this narrower account. The court was careful to say that it was not deciding a generative AI case: Ross's system "was not a generative AI," slip op. at 5, and the court distinguished "the AI models in *Bartz* and *In re: OpenAI*," which can "generate original expression," *id.* at 17 n.7. Its decision rested on head-to-head competition: Ross "aspired to be a direct competitor by using [Westlaw's headnotes] for a highly similar purpose," *id.* at 2, and "aim[ed] to replace Westlaw," *id.* at 19. The court treated that competition as making the use "minimally transformative, at best," *id.* at 17, and weighed it heavily under the fourth factor, *id.* at 24–27. The court also found cognizable harm to a "rapidly developing" market for licensing headnotes as AI training data. *Id.* at 26. But that market involved a small, identified corpus, a defendant that copied it to build a direct competitor, and a plaintiff already using the same

material to train its own products; it bears no resemblance to the internet-scale licensing problem described in Part II.B.3. With respect, courts would do better to analyze economic competition under the fourth factor rather than the first. But on either approach, *Ross* is a case about a product built from one specific source in order to replace that same source, not about generalized competition from new works.

This Court has already dismissed the Times and Daily News plaintiffs' hot-news misappropriation claims as preempted. *New York Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283, 319–22 (S.D.N.Y. 2025). That does not mean, however, that considerations analogous to a well-pleaded unfair competition complaint have no place in the fourth factor inquiry. But general speculation that authors and publishers as a class may be disadvantaged by the arrival of generative AI is nowhere close to either direct or indirect expressive substitution.

E. "Lawful access" may be relevant to the fourth factor, in appropriate circumstances.

Copyright owners have no right to charge for transformative uses, but they do have a right to charge for access to their works. A commercial user that bypasses the market for access without a compelling reason arguably undermines the economic incentives that copyright is designed to create. *Fairness*, 92 Fordham L. Rev. at 1917–18. That, rather than any theory of separate uses, is the core of Judge Alsup's objection in *Bartz*: Anthropic "ha[d] many places from which" it could have purchased books "but it preferred to steal them," 787 F. Supp. 3d at 1015; it built its library "without paying for it," *id.* at 1026; and, unlike Texaco, it "never paid for the central library copies stolen off the internet," *id.* at 1027.

This is not a clean-hands requirement. Fair use does not categorically require that the defendant begin with an authorized copy. *Kadrey*, 788 F. Supp. 3d at 1046 (the Federal Circuit's suggestion to the contrary in *Atari* "overread the cases on which it relied"); Leval, 103 Harv. L.

Rev. at 1126–28. Courts have found fair use where the defendant worked from unauthorized copies. *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 77, 83 (2d Cir. 2014) (recording obtained "without authorization"); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1164 n.8 (9th Cir. 2007) (access to infringing websites).

It would be unwise to elevate lawful access to a per se rule, however, even for commercial defendants. *Fairness*, 92 Fordham L. Rev. at 1918. The relevant question is whether, by obtaining books as it did, OpenAI bypassed a market for access that was actually available to it in a way that was likely to cause harm. OpenAI states that it used material from Library Genesis only for GPT-3 and GPT-3.5 and for no other model at issue in this case, ECF No. 1738 ¶¶ 10–11, although the plaintiffs dispute the point, ECF No. 2195 at 21–22. At the relevant time, publishers did not offer collections of digital books for computational use, and consumer e-book licenses likely forbade it. So OpenAI's alternative was to buy print copies by the million and destructively scan them as Anthropic eventually did. *Bartz* held that format change to be fair use, 787 F. Supp. 3d at 1019, 1023, and yet that practice, too, now draws objection. *See, e.g.*, Ashley Gold, *Exclusive: FTC Urged to Investigate AI Firms for Destroying Books*, AXIOS (Aug. 21, 2026), https://www.axios.com/2026/08/21/ftc-ai-companies-book-destruction-investigate.

A defendant cannot be faulted for bypassing a market that did not exist. But the Court will have to decide whether buying and scanning print copies was a reasonable alternative in all the circumstances.

CONCLUSION

Accordingly, amicus respectfully submits that the Court should hold that training the models at issue on the plaintiffs' works was transformative; and determine whether OpenAI's acquisition of books from shadow libraries is an issue that survives summary judgment on this record.

Dated: October 29, 2026

Respectfully submitted,

By: */s/ Ethan Jacobs*
Ethan Jacobs
Ethan Jacobs Law Corporation
100 Pine Street, Suite 1250
San Francisco, CA 94111
(415) 275-0845
ethan@ejacobslaw.com

Counsel for Amicus Curiae Professor Matthew Sag

CERTIFICATE OF COMPLIANCE

Pursuant to Local Civil Rule 7.1(c), I certify that this brief contains 5,951 words, excluding the caption, table of contents, table of authorities, signature block, and this certificate, as counted by the word-count function of Microsoft Word, and that it complies with the 6,000-word limit set by the Court's Order of August 7, 2026 (ECF No. 1663). The brief is set in 12-point Times New Roman type, double-spaced, with one-inch margins, in compliance with Local Civil Rule 7.1(b).

Dated: October 29, 2026

By: *<u>/s/ Ethan Jacobs</u>*
Ethan Jacobs
Ethan Jacobs Law Corporation
100 Pine Street, Suite 1250
San Francisco, CA 94111
(415) 275-0845
ethan@ejacobslaw.com

Counsel for Amicus Curiae Professor Matthew Sag