UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION THIS DOCUMENT RELATES TO: ALL CASES	25-md-3143 (SHS) (OTW)

BRIEF OF PUBLIC KNOWLEDGE AS AMICUS CURIAE IN SUPPORT OF DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT IN PART

John Bergmayer
Meredith Rose
Nicholas Garcia
PUBLIC KNOWLEDGE
1818 N Street NW, Suite 410
Washington, DC 20036
(202) 861-0020
john@publicknowledge.org
Counsel for Proposed Amicus Curiae Public Knowledge

CORPORATE DISCLOSURE STATEMENT

Public Knowledge is a nonprofit corporation. It has no parent corporation and does not issue stock. No publicly held corporation owns 10% or more of Public Knowledge.

TABLE OF CONTENTS

CORPORATE DISCLOSURE STATEMENT i
INTEREST OF AMICUS CURIAE 1
INTRODUCTION 1
ARGUMENT 4
I. Training Must Be Evaluated as a Separate and Distinct Use 4
II. Copying Protectable Expression in the Course of Training Large Language Models Is a Fair Use 7
A. Copying Protectable Expression for Training Is a Fair Use 7
B. Fair Use Permits Reasonably Necessary Copying to Access Unprotected Information, Which Includes Facts About Creative Works 11
1. Facts About a Creative Work Are Not Protectable 12
2. Copyright Law Allows for Copying of Protected Expression to Access Unprotectable Facts 14
CONCLUSION 19
CERTIFICATE OF COMPLIANCE 20
CERTIFICATE OF SERVICE 21

TABLE OF AUTHORITIES

Cases

Am. Soc'y for Testing & Materials v. Public.Resource.Org, Inc., 82 F.4th 1262 (D.C. Cir. 2023) 3
Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023) 5, 6
Arica Inst., Inc. v. Palmer, 970 F.2d 1067 (2d Cir. 1992) 10
Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832 (Fed. Cir. 1992) 14
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) 3, 11, 16
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) 11
Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) 9
Chapman v. Maraj, No. 2:18-cv-09088-VAP-SSx, 2020 U.S. Dist. LEXIS 198684 (C.D. Cal. Sept. 16, 2020) 5
Eldred v. Ashcroft, 537 U.S. 186 (2003) 3, 12
Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) 12, 13
Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018) 6
Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539 (1985) 12, 13
Hoehling v. Universal City Studios, Inc., 618 F.2d 972 (2d Cir. 1980) 13
Kadrey v. Meta Platforms, Inc., 788 F. Supp. 3d 1026 (N.D. Cal. 2025) 9, 10, 16
Matthew Bender & Co. v. West Publ'g Co., 158 F.3d 693 (2d Cir. 1998) 15
Mazer v. Stein, 347 U.S. 201 (1954) 13
N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. Apr. 4, 2025), ECF No. 514 ..... 7, 8, 11
Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) 2
Romanova v. Amilus Inc., 138 F.4th 104 (2d Cir. 2025) 10, 17
Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992) 10, 14, 15, 16
Sony Comput. Entm't, Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000) ..... passim
Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P., 756 F.3d 73 (2d Cir. 2014) 15
Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc., No. 25-2153, 2026 WL 2932557 (3d Cir. Sept. 29, 2026) 8, 17
Wright v. Warner Books, Inc., 953 F.2d 731 (2d Cir. 1991) 10

Statutes

17 U.S.C. § 102(b) 11
17 U.S.C. § 107 5
17 U.S.C. § 1202(b)(3) 11

Other Authorities

Class Plaintiffs' Memorandum of Law in Opposition to OpenAI Defendants' Motion for Summary Judgment, In re OpenAI, Inc., Copyright Infringement Litigation, No. 25-md-03143 (S.D.N.Y. Oct. 24, 2026), ECF No. 2271 12

Jean-Baptiste Michel et al., Quantitative Analysis of Culture Using Millions of Digitized Books, 331 Science 176 (2011) 12, 17

Kristina Gulordava et al., Colorless Green Recurrent Networks Dream Hierarchically, in 1 Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1195 (Marilyn Walker et al. eds., 2018) 17

News Plaintiffs' Combined Summary Judgment Brief, In re OpenAI, Inc., Copyright Infringement Litigation, No. 25-md-03143 (S.D.N.Y. Oct. 16, 2026), ECF No. 2108 4

News Plaintiffs' Omnibus Memorandum of Law in Opposition to Defendants' Motions for Summary Judgment, In re OpenAI, Inc., Copyright Infringement Litigation, No. 25-md-03143 (S.D.N.Y. Oct. 23, 2026), ECF No. 2118 12

Statement of Interest of the United States, In re OpenAI, Inc., Copyright Infringement Litigation, No. 25-md-03143 (S.D.N.Y. Sept. 1, 2026), ECF No. 1682 2

U.S. Copyright Off., Copyright and Artificial Intelligence, Part 3: Generative AI Training (pre-publication version May 2025)..... 7

Yoshua Bengio et al., A Neural Probabilistic Language Model, 3 J. Mach. Learning Rsch. 1137 (2003)..... 17

INTEREST OF AMICUS CURIAE

Public Knowledge is a nonprofit public-interest organization dedicated to preserving an open internet and the public's access to knowledge, promoting creativity through balanced intellectual property policies, and protecting consumers' ability to use innovative technology lawfully. Public Knowledge seeks to protect the legitimate interests of authors while preserving the public's ability to access information, and submits this brief because rules governing AI training will affect independent researchers, open-source developers, libraries, and internet users as well as the parties.[†]

INTRODUCTION

A free, open, democratic, and well-informed society requires an informed citizenry with access to news and information of all kinds. That includes books on every topic and of every genre, journalism, as well as news analysis, commentary, and opinion pieces. The vast digital commons of the open internet has put more information at people's fingertips than at any time in history. Search engines and social media have reshaped how it is distributed and monetized. The public has an interest both in sustaining the creation of these works and in being able to access and learn from them.

The sheer amount of content and information available can be overwhelming. One promise of artificial intelligence (AI) is to help tame the unprecedented volume of information now available to us. It can help users find what they need and understand what they find in ways that complement existing tools such as search. It can also help new competitors challenge incumbent gatekeepers, increase access to knowledge, and create new opportunities for authors and

[†] No party or party's counsel authored this brief in whole or in part. No person other than Public Knowledge or its counsel contributed money intended to fund the preparation or submission of this brief.

publishers. But whether new technology produces public benefits depends on who controls this technology, how they deploy it, and how it can be used. Tools that make information easier to access can also increase the power of gatekeeping intermediaries.

For years, publishers have been adapting to the changing demands of the digital landscape, including the rise of powerful tech gatekeepers. Many publishers have even shut down, been acquired, or merged with other entities—which raises its own concerns about media concentration and the loss of independent voices. Concerns that AI could deepen those dependencies, or further industry consolidation, deserve serious attention as a matter of public policy.

But copyright is not a general remedy for technological disruption or a means for reallocating the gains from a new technology. Preserving a robust understanding of fair use is itself part of protecting the public's ability to benefit from new technology and ensuring a healthy information ecosystem. A finding that model training is not fair use, by contrast, could favor the largest technology and media companies. As the United States explains, "only the largest technology companies might have the capital necessary to pay licensing fees," while "such licensing fees would disproportionately benefit legacy media outlets." Statement of Interest of the United States at 4, In re OpenAI, Inc., Copyright Infringement Litigation, No. 25-md-03143 (S.D.N.Y. Sept. 1, 2026), ECF No. 1682. At the same time, fair use does not excuse anticompetitive behavior or other violations of law by AI developers, even if the specific acts of copying are lawful. Nor does it foreclose policies designed to ensure that the benefits of new technology are more widely shared.

There is precedent for this. Fair use has protected tools that help people navigate the vast collections of information the digital world has brought us, including image search and full-text book search. *See Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1164–68 (9th Cir. 2007);

Authors Guild v. Google, Inc., 804 F.3d 202, 216–17, 225 (2d Cir. 2015) ("*Google Books*"). It has also protected libraries' provision of accessible digital books to readers with print disabilities. *See Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 101–03 (2d Cir. 2014). And it has protected free online publication of technical standards incorporated into law, recognizing the "substantial public benefits of free and easy access to the law." *Am. Soc'y for Testing & Materials v. Public.Resource.Org, Inc.*, 82 F.4th 1262, 1271–72 (D.C. Cir. 2023). Fair use will be equally beneficial in ensuring that AI's benefits accrue to the public at large, instead of being controlled by a privileged few.

Fair use is integral to copyright's purpose and design. Copyright's purpose is to promote progress and expand public knowledge, both by rewarding authors and by preserving the ability of others to learn from and build upon their works. Its boundaries are integral to that purpose. As the Supreme Court explained in *Eldred v. Ashcroft*, copyright contains "built-in First Amendment accommodations": the idea/expression dichotomy leaves facts and ideas available to the public, and fair use preserves permissionless copying of protected expression in appropriate circumstances. 537 U.S. 186, 219–20 (2003).

The Court must identify the specific act of copying or other challenged use. AI development can involve acquiring works; creating training datasets; copying works during training; retrieving material to ground responses; running trained models in response to prompts (inference); and generating outputs, which may involve reproduction of protected expression. These activities may differ in terms of what is copied, why it is being copied, and the effect of this on the market for works. Fair use turns on the specific use challenged.

The vertically integrated products offered by OpenAI and Microsoft are not the only approach to AI. Many AI researchers and developers in the academic, open-source, and public

interest sectors do not operate in the same vertically integrated fashion as OpenAI and Microsoft. Many of these actors do not directly control the downstream use of the data they collect or the models they train. For example, the record in this case indicates that OpenAI benefitted from access to resources like Common Crawl. *See* News Plaintiffs' Combined Summary Judgment Brief at 9, *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 25-md-03143 (S.D.N.Y. Oct. 16, 2026), ECF No. 2108. There are open-source and open-weight model developers, researchers, auditors, and analysts who rely on the ability to inspect training data, train and release models freely, directly access and modify model weights, and replicate the same kind of evaluations used by the AI companies. Rules developed around the defendants' integrated products could affect what they can do, without taking into account how they increase the benefit the public gets from new technologies and access to knowledge.

Public Knowledge therefore urges the Court to consider the pivotal issue of AI training separately from data acquisition and AI's outputs. To be clear, Public Knowledge believes that fair use or other doctrines will often protect other uses involved in AI development and deployment. But those uses each present distinct questions. Thus, this brief principally addresses the intermediate copying of works that is a necessary part of training AI models. Recognizing fair use for this copying would preserve opportunities for a broad range of researchers and developers to build tools that benefit the public, while leaving other uses, if challenged, subject to their own distinct analysis.

ARGUMENT

I. Training Must Be Evaluated as a Separate and Distinct Use

Copying works for AI training must be evaluated as a distinct use, separate from acquiring the training materials or producing and distributing particular outputs. Fair use turns on the specific

use challenged. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 533–34 (2023). Training that qualifies as fair use does not lose that status because of a later infringing use.[1] Nor does fair use in training excuse later infringement: each challenged use requires its own analysis under the four statutory factors. *See* 17 U.S.C. § 107; *Warhol*, 598 U.S. at 533–34.

Warhol illustrates why the creation of a secondary work and subsequent uses of that work may require distinct fair-use analyses. Importantly, each of those uses must, per the Court, be analyzed separately. *See Warhol*, 598 U.S. at 533–34. This is a close fit for AI, where distinct fair use questions may arise at multiple stages, including training and the production or distribution of particular outputs.

In *Warhol*, magazine Vanity Fair obtained a license for photographer Lynn Goldsmith's portrait of the artist Prince. The magazine then approached Andy Warhol to create a silkscreen print based upon the photograph. Unbeknownst to either Vanity Fair or Goldsmith, Warhol made not one print, but a series of sixteen works: fourteen silkscreen prints and two pencil drawings. After Prince's passing, the Andy Warhol Foundation and Vanity Fair's parent company (Condé Nast) negotiated a deal to feature one of the additional prints on the magazine's cover. *Id.* at 516–20.

The fair use analysis could focus on whether using Goldsmith's photograph in the creation of the fifteen unlicensed works was fair. However, the Supreme Court distinguished the creation of the unlicensed works, and the "multiple ways" *the secondary works incorporating the photograph* were used. *Id.* at 533–34. The Court listed numerous (real and hypothetical) "uses" of the secondary works—unlicensed creation, commercial publication of an unlicensed print,

[1] *See Chapman v. Maraj*, No. 2:18-cv-09088-VAP-SSx, 2020 U.S. Dist. LEXIS 198684, at *33–34 (C.D. Cal. Sept. 16, 2020) (holding creation of a recording for artistic experimentation and license approval fair use while leaving the separate distribution claim unresolved).

commercial licensing of a licensed print, and display in a gallery setting, among others—and took great pains to emphasize that each of these required a separate analysis: "[t]he same copying may be fair when used for one purpose but not another." *Warhol*, 598 U.S. at 533–34.[2]

In other words, "fair use" is not an abstract status that attaches to a work if the circumstances of its initial creation are right. Instead, it is an assessment of whether or not each challenged use, including subsequent re-use of the secondary work, is "fair"—including commercially—vis-à-vis the source material. Courts have always been able to ask, was the original work used fairly *in the creation* of the secondary work? However, courts can also ask, is *this particular use of the secondary work* fair?

Importantly, a "no" on the second question does not require a "no" on the first. The Court held that the first factor did not favor "the specific use" at the center of its analysis ("AWF's licensing of *Orange Prince* to Condé Nast"). *Warhol*, 598 U.S. at 525–26, 550–51. Nevertheless, the Court reiterated that it "expresses no opinion as to the **creation**, display, or sale of any of the original Prince Series works." *Id.* at 534 (emphasis added). Nor is the first-factor finding dispositive for any of the *other* uses made between the point of creation and the disputed use. As Justice Gorsuch explained in concurrence:

> Last but hardly least, while our interpretation of the first fair-use factor does not favor the Foundation in this case, it may in others. If, for example, the Foundation had sought to display Mr. Warhol's image of Prince in a nonprofit museum or a for-profit book commenting on 20th-century art, the purpose and character of that use might well point to fair use.

Id. at 557–58 (Gorsuch, J., concurring). While evidence of model outputs may bear on the initial training's purpose or market effects, this does not mean that all uses should be merged into a single fair-use inquiry. The distinction is between functionally different uses, not every technical copy

[2] *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 176 (2d Cir. 2018) (analyzing distinct functions separately; search was unchallenged on appeal, while viewing was held not to be fair use).

made to accomplish one use. *Cf.* U.S. Copyright Off., *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 36–37 (pre-publication version May 2025) (recognizing that different uses require separate consideration while urging evaluation in the context of the overall use).

The Court must identify the specific act of copying or other challenged use. The analysis must distinguish data collection, training, retention of copies in model weights, actual model outputs, and even hypothetical market impacts of hypothetical future model outputs. These distinctions matter.

II. Copying Protectable Expression in the Course of Training Large Language Models Is a Fair Use

For training copies, the relevant use is analytical training. *See Sony Comput. Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 602, 605 (9th Cir. 2000). Copying of human-authored works during the training phase of AI development is protected by fair use, informed by the idea/expression dichotomy. First, the use of the protected original works is highly transformative. Second, the purpose of that copying is to access the works to calculate statistical patterns and relationships embodied in the work.

A. Copying Protectable Expression for Training Is a Fair Use

The relevant copying is the intermediate copies made of the dataset, including the "tokens" derived from the original works used during training. As the Court summarized plaintiffs' allegations, the training process requires a developer to

> (1) store copies of the training articles in computer memory, (2) encode the information from the training dataset in a numerical format, (3) provide a portion

> of the article to the model, and then (4) adjust the parameters of the model so that the model accurately predicts the next word in the article.

N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195, slip op. at 8 (S.D.N.Y. Apr. 4, 2025), ECF No. 514 (alterations and internal quotation marks omitted). Each of the four fair use factors weighs in favor of this use being fair.

Purpose of the use. Training is a highly transformative activity: it analyzes works to calculate statistical relationships among tokens. It does not provide copies of those works, or of protectable expression, to readers as a substitute for the original works in the marketplace. The model weights that derive from this analysis are distinct from both the original training data, and from outputs generated in response to users. As in *Google Books*, making a copy to extract information about the underlying works serves a transformative purpose. *Google Books*, 804 F.3d at 216–17. Similarly, in *Bartz*, the court held that "the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative." *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 2025 WL 1741691, at *8 (N.D. Cal. June 23, 2025).[3]

Nature of the copyrighted work. This factor weighs in favor of fair use because AI training involves copying works necessary for its transformative purpose. Even creative works contain unprotected factual information about, or embedded in, their expression; and even factual accounts reflect an author's expressive choices. Copyright protects these creative choices, but it does not protect facts, and it does not prevent the copying and analysis of expression necessary to obtain those facts. *See Google Books*, 804 F.3d at 220. As the Second Circuit explained, "If one (or all)

[3] *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.* does not establish that all AI training is nontransformative. No. 25-2153, 2026 WL 2932557, at *6–9 & n.7 (3d Cir. Sept. 29, 2026). There, copying headnotes served the same purpose as Thomson Reuters' Westlaw: a legal-research service. The court expressly distinguished this use from generative models capable of producing original expression, including those at issue here. *Id.* at *6 n.7.

of the plaintiff works were fiction, we do not think that would change in any way our appraisal." *Google Books*, 804 F.3d at 220.

Amount and substantiality of the portion used. Copying entire works can be reasonably necessary to access unprotected facts. As *Google Books* found, "Complete 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." *Google Books*, 804 F.3d at 221. Applying that principle to AI training, *Kadrey* concluded that "this factor favors Meta, even though it copied the plaintiffs' books in their entirety." *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1050 (N.D. Cal. 2025) (amount used reasonable in relation to the training purpose). Similarly, *Connectix* distinguished "the necessity of the method, i.e., disassembly, not the necessity of the number of times that method was applied." 203 F.3d at 605.

Effect of the use on the market for the original work. Factor four examines whether the challenged use substitutes for the original work. *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 591–93 (1994). When a use is transformative, "market substitution is at least less certain, and market harm may not be so readily inferred." *Id.* at 591. Copying works for analytical training does not, by itself, supply readers with substitutes for those works.

Plaintiffs argue that AI outputs can still cause harm ultimately traceable to unlawful training copies. But the Second Circuit has already addressed this argument, distinguishing harm

from competition from harm caused directly by the use of protected expression in downstream works. In *Arica*, the court explained:

> Nonetheless, we believe that the fourth factor favors defendants. As we recently stated in *Wright*, the relevant market effect is that which stems from defendant's use of plaintiff's "expression," not that which stems from defendant's work as a whole. *Wright*, 953 F.2d at 739. Defendant, after all, is perfectly entitled to create a competing work. The copyright statute simply constrains her from using plaintiff's original expression in doing so.

Arica Inst., Inc. v. Palmer, 970 F.2d 1067, 1078 (2d Cir. 1992). (*Kadrey*'s contrary reasoning does not explain why the scale of competition should expand the interests copyright protects. *See* 788 F. Supp. 3d at 1054–55.)

In *Sega*, Accolade copied Sega's programs during development to create independently designed games that competed with Sega's games. In finding that the fourth factor favored fair use, the court distinguished supplanting the copyrighted work from "a use which simply enables the copier to enter the market for works of the same type as the copied work." *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1523–24 (9th Cir. 1992). Likewise, *Connectix* recognized that a product developed through intermediate copying could reduce Sony's sales, but held that "some economic loss by Sony as a result of this competition does not compel a finding of no fair use." 203 F.3d at 607.

The Court should therefore reject a "market dilution" theory based on AI models (after training) facilitating the creation of new, noninfringing works. An author's copyright does not reserve the market for similar works created by others. As the Second Circuit explained, fair use considers whether the copy will "supersede or supplant the original, serving as a market substitute for the protected expression of the original." *Romanova v. Amilus Inc.*, 138 F.4th 104, 117 n.9 (2d Cir. 2025). In the AI context, *Bartz* similarly rejected the notion that an increase in competing works constituted cognizable market harm. *Bartz*, 2025 WL 1741691, at *17. As the court

explained, "The Act seeks to advance original works of authorship, not to protect authors against competition." *Id.* Accordingly, the fourth factor supports a finding of fair use.

The fact that some publishers might want to charge for (or deny) this use does not expand the scope of their rights. A rightsholder cannot obtain an exclusive right over unprotected information merely by offering licenses to it. *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (2d Cir. 2006). Because the challenged use occupied a transformative market, the court held that "BGA does not suffer market harm due to the loss of license fees." *Id.* at 615. *HathiTrust* accordingly rejected lost search-license fees, while *Google Books* rejected an exclusive right to supply information about works. *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 100 (2d Cir. 2014); *Google Books*, 804 F.3d at 225–26.

Uses that incorporate safeguards that prevent substitutive downstream reproductions present an even easier case. Whether the models in this case have such safeguards must be assessed on the record. In addressing the pleaded DMCA section 1202(b)(3) claim, the Court stated that "the regurgitations do not constitute substantial or entire reproductions of plaintiffs' works." *N.Y. Times*, slip op. at 29 (alterations and internal quotation marks omitted). Where "at best and after a large commitment of manpower" a reader is only able to obtain "discontinuous, tiny fragments" there is no "access to effectively competing substitutes." *Google Books*, 804 F.3d at 222–24.

B. Fair Use Permits Reasonably Necessary Copying to Access Unprotected Information, Which Includes Facts About Creative Works

As the Copyright Act states: "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." 17 U.S.C. § 102(b). Plaintiffs' argument—that intermediate copying in order to extract unprotectable facts about a work is not permitted if that copying could be used to enable

future production of infringing (or even merely competing) works—would not only run contrary to decades of caselaw, but would functionally erase the boundary between idea and expression.[4]

1. Facts About a Creative Work Are Not Protectable

This case turns on facts concerning the works themselves: word frequencies and corpus-wide linguistic relationships.[5] For example, a book may report a fact about whales; the frequency with which it uses the word "whale" is instead a fact about the work and its use of language. While copyright protects the works' expression, it does not give authors a monopoly over learning or using objective information about their works. Learning these patterns is distinct from reproducing a work's protected sequence of words.

The Supreme Court has repeatedly underscored the importance of this distinction. Under copyright law, authors have rights only to their original expression; they do not have control over the facts and ideas that may be learned from their works. *See, e.g., Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 547 (1985) (internal citations omitted) ("[N]o author may copyright facts or ideas. The copyright is limited to those aspects of the work—termed 'expression'—that display the stamp of the author's originality."); *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003) (citing *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349–50 (1991)) ("First, [copyright law] distinguishes between ideas and expression and makes only the latter eligible for copyright protection. . . . [E]very idea, theory, and fact in a copyrighted work becomes

---

[4] *See* News Plaintiffs' Omnibus Memorandum of Law in Opposition to Defendants' Motions for Summary Judgment at 8–16, 20–31, *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 25-md-03143 (S.D.N.Y. Oct. 23, 2026), ECF No. 2118; Class Plaintiffs' Memorandum of Law in Opposition to OpenAI Defendants' Motion for Summary Judgment at 18–39, *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 25-md-03143 (S.D.N.Y. Oct. 24, 2026), ECF No. 2271 (arguing that training and competing outputs cause cognizable harm).

[5] *See, e.g.,* Jean-Baptiste Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 331 Science 176 (2011), https://doi.org/10.1126/science.1199644 (analyzing word and phrase frequencies across a corpus of digitized books).

instantly available for public exploitation at the moment of publication."); *Mazer v. Stein*, 347 U.S. 201, 217 (1954) ("Unlike a patent, a copyright gives no exclusive right to the art disclosed; protection is given only to the expression of the idea—not the idea itself.") (footnote omitted). Plaintiffs cannot circumvent this rule by pointing out that defendants learned the facts from (or about) plaintiffs' protected expression:

> The most fundamental axiom of copyright law is that "[n]o author may copyright his ideas or the facts he narrates." . . . [F]acts do not owe their origin to an act of authorship. The distinction is one between creation and discovery: The first person to find and report a particular fact has not created the fact; he or she has merely discovered its existence. . . . Others may copy the underlying facts from the publication, but not the precise words used to present them. . . . [C]opyright assures authors the right to their original expression, but encourages others to build freely upon the ideas and information conveyed by a work. This principle, known as the idea/expression or fact/expression dichotomy, applies to all works of authorship.

Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 344–45, 347–50 (1991) (quoting *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 556 (1985)) (citation omitted). *See also Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972, 974 (2d Cir. 1980) ("the cause of knowledge is best served when [it] is the common property of all, and each generation remains free to draw upon the discoveries and insights of the past.").

As in Google Books, "the secondary use transformatively provides valuable information about the original, rather than replicating protected expression in a manner that provides a meaningful substitute for the original," 804 F.3d at 220. The expressive works in this case are replicated so that the training program can extract and analyze data *about* them—not simply to replicate them in a manner that "provides a meaningful substitute for the original," *id.*, by "bodily appropriat[ing] the expression of another," *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972, 980 (2d Cir. 1980), and placing it into the marketplace.

2. Copyright Law Allows for Copying of Protected Expression to Access Unprotectable Facts

Because facts in and about a copyrighted work are not protected, courts have repeatedly held that copyright law does not stand in the way of accessing those facts. "The Copyright Act permits an individual in rightful possession of a copy of a work to undertake necessary efforts to understand the work's ideas, processes, and methods of operation." *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832, 842 (Fed. Cir. 1992).[6] This includes making new copies of works, as necessary, for analytical purposes:

> [H]umans often cannot gain access to the unprotected ideas and functional concepts contained in object code without disassembling that code—i.e., making copies. . . . If disassembly of copyrighted object code is per se an unfair use, the owner of the copyright gains a de facto monopoly over the functional aspects of his work—aspects that were expressly denied copyright protection by Congress. . . . [W]here disassembly is the only way to gain access to the ideas and functional elements embodied in a copyrighted computer program and where there is a legitimate reason for seeking such access, disassembly is a fair use of the copyrighted work, as a matter of law.

Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1525–28 (9th Cir. 1992).

Connectix applied the same rule, and further explained why protected expression must sometimes be copied in order to reach unprotected material.

[6] Importantly, that limit does not excuse unlawful acquisition. *Bartz* distinguished training copies from a pirated general-purpose library, explaining that a person intending fair use is not entitled "to steal a copy," and that library copying was "a different use altogether." *Bartz*, 2025 WL 1741691, at *18. *Atari* likewise approved only "[r]everse engineering, untainted by the purloined copy . . . and necessary to understand 10NES." *Atari*, 975 F.2d at 843–44. This argument addresses copying for training, not whether particular source copies were lawfully acquired.

> The object code of a program may be copyrighted as expression . . . but it also contains ideas and performs functions that are not entitled to copyright protection. . . . Object code cannot, however, be read by humans. The unprotected ideas and functions of the code therefore are frequently undiscoverable in the absence of investigation and translation that may require copying the copyrighted material.

Sony Comput. Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 602 (9th Cir. 2000). The Second Circuit applied similar reasoning to its ruling in *Google Books*:

> [T]he purpose of Google's copying of the original copyrighted books is to make available significant information *about* those books. . . . The copyright resulting from the Plaintiffs' authorship of their works does not include an exclusive right to furnish the kind of information about the works that Google's programs provide to the public.

Google Books, 804 F.3d at 217, 225; *cf. Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 84–85 (2d Cir. 2014) (expressive choices such as tone, cadence, and hesitation convey information); *see also Matthew Bender & Co. v. West Publ'g Co.*, 158 F.3d 693, 699–700 (2d Cir. 1998) (holding that mechanically generated volume and page numbers convey unprotected information).

In *Connectix*, the court found that these temporary copies being made were fair uses. Instead of focusing on the magnitude of copying, the court inquired whether it was reasonably necessary, and found that "Connectix's intermediate copying and use of Sony's copyrighted BIOS was a fair use for the purpose of gaining access to the unprotected elements of Sony's software." *Connectix*, 203 F.3d at 602. The court continued,

> The "necessity" we addressed in Sega was the necessity of the method, i.e., disassembly, not the necessity of the number of times that method was applied. . . . Such an approach would erect an artificial hurdle in the way of the public's access to the ideas contained within copyrighted software programs. These are "aspects that were expressly denied copyright protection by Congress." We decline to erect such a barrier in this case.

Connectix, 203 F.3d at 605 (citation omitted). Thus, for AI training, the relevant inquiry is whether the scope of copying was reasonably necessary to access the unprotectable

information (to use the language of *Connectix*, the "necessity of the method"). 203 F.3d at 605.[7]

Courts have consistently allowed full-text copying of large corpuses in order to achieve transformative purposes. In *HathiTrust*, complete digital copies were necessary for a transformative purpose—namely, non-substitutive, full-text search:

> It is not disputed that, in order to perform a full-text search of books, the Libraries must first create digital copies of the entire books. . . . [T]he creation of a full-text searchable database is a quintessentially transformative use. . . . [T]he result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn. . . . Because it was reasonably necessary for the HDL to make use of the entirety of the works in order to enable the full-text search function, we do not believe the copying was excessive.

Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97–98 (2d Cir. 2014).

Google Books likewise measured the amount copied against the *analytical purpose* for the copying:

[7] Notably, *Kadrey* reasoned that training differs from software reverse engineering because it depends on books' creative expression, including word choice, order, grammar, and syntax. *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1049–50 (N.D. Cal. 2025). This distinction is, in reality, illusory. In *Sega* and *Connectix*, the defendants copied entire copyrighted programs, including their expressive elements. *See Sega*, 977 F.2d at 1526; *Connectix*, 203 F.3d at 605–06. The compatibility information they sought was information about, and that depended on, protected expression. Information learned from expressive choices is not necessarily itself protected expression. *Kadrey* in the end fails to distinguish software reverse engineering from AI model training; disqualifying analytical copying because it depends on expressive choices would foreclose the very reverse engineering those cases permit.

> Notwithstanding that the libraries had downloaded and stored complete digital copies of entire books, we noted that such copying was essential to permit searchers to identify and locate the books in which words or phrases of interest to them appeared. . . . [N]ot only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose. If Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times).

Google Books, 804 F.3d at 217, 221; *see also Romanova v. Amilus Inc.*, 138 F.4th 104, 115–17 & n.9 (2d Cir. 2025) (distinguishing making copies for informational uses from substituting for a work in the marketplace by substantially reproducing it).

The unprotectable information needed for LLM training is contained across many works, and is derived by examining a sufficiently large corpus of texts.[8] No particular work need be indispensable to corpus analysis: researchers need not use *The Great Gatsby* specifically to teach a model noun-verb agreement or other facts about language use. What matters is whether copying works across a sufficiently large corpus is reasonably necessary for that analysis, not whether each work is irreplaceable. *ROSS*'s rejection of copying for mere ease does not eliminate the need to evaluate the reasonable necessity of the method for a genuinely different purpose. *See ROSS*, 2026 WL 2932557, at *7.

Courts have already grappled with this precise issue. The court in *Bartz*, 2025 WL 1741691, assumed that Anthropic could have trained using different books or no books, but found a compelling justification for using the specific books it selected. *Id*. at *15–16. But as the court explained, "'reasonably necessary' does not mean 'strictly necessary.'" *Id.* at *16. Given the

[8] For examples of corpus-based analysis of language, *see* Jean-Baptiste Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 331 Science 176 (2011) (analyzing linguistic and cultural patterns across digitized books); Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. Mach. Learning Rsch. 1137, 1138–39 (2003) (learning word-sequence probabilities from text corpora); Kristina Gulordava et al., *Colorless Green Recurrent Networks Dream Hierarchically*, in 1 Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1195, 1198–1200 (Marilyn Walker et al. eds., 2018) (learning grammatical agreement from Wikipedia corpora).

volume of text required, "using any one work for actually training LLMs was about as reasonable as the next." *Id.* The availability of alternative works, therefore, did not overcome a finding of fair use. *Bartz* also explained why model training is transformative.

> Anthropic used copies of Authors' copyrighted works to iteratively map statistical relationships between every text-fragment and every sequence of text-fragments so that a completed LLM could receive new text inputs and return new text outputs . . . . In short, the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative. . . . If this training process reasonably required making copies within the LLM or otherwise, those copies were engaged in a transformative use.

Bartz, 2025 WL 1741691, at *7–8.

Accordingly, reasonably necessary transformative copying supports fair use even when the entire work must be copied.

CONCLUSION

For the foregoing reasons, the Court should hold that the challenged training use is a fair use.

Dated: October 29, 2026
Respectfully submitted,
/s/ John Bergmayer
John Bergmayer
PUBLIC KNOWLEDGE
1818 N Street NW, Suite 410
Washington, DC 20036
(202) 861-0020
john@publicknowledge.org
Counsel for Proposed Amicus Curiae Public Knowledge

CERTIFICATE OF COMPLIANCE

Pursuant to Local Civil Rule 7.1(c) and the Court's orders setting a 6,000-word limit for amicus briefs, I certify that this brief contains 5,877 words, including footnotes, the corporate disclosure statement, and the certificate of service, and excluding the caption, tables of contents and authorities, signature blocks, and this certificate of compliance. I relied on Microsoft Word's word-count feature in preparing this certificate.

Dated: October 29, 2026

/s/ John Bergmayer
John Bergmayer

CERTIFICATE OF SERVICE

I certify that on October 29, 2026, I caused this brief to be filed electronically through the Court's CM/ECF system. Notice will be sent by operation of the Court's electronic filing system to all registered counsel of record.

Dated: October 29, 2026
/s/ John Bergmayer
John Bergmayer