UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: 1:23-cv-08292 (SHS) (OTW) 1:23-cv-10211 (SHS) (OTW) 1:25-cv-03482 (SHS) (OTW) 1:23-cv-11195 (SHS) (OTW) 1:24-cv-01515 (SHS) (OTW) 1:24-cv-03285 (SHS) (OTW) 1:24-cv-04872 (SHS) (OTW) 1:25-cv-04315 (SHS) (OTW)	25-md-3143 (SHS) (OTW)

BRIEF OF AMICI CURIAE ASSOCIATION OF AMERICAN PUBLISHERS, COPYRIGHT CLEARANCE CENTER, INDEPENDENT BOOK PUBLISHERS ASSOCIATION, AND INTERNATIONAL ASSOCIATION OF SCIENTIFIC, TECHNICAL & MEDICAL PUBLISHERS IN SUPPORT OF CLASS PLAINTIFFS' AND NEWS PLAINTIFFS' MOTIONS FOR SUMMARY JUDGMENT

TABLE OF CONTENTS

TABLE OF AUTHORITIES ii

STATEMENT OF INTERESTS OF AMICI CURIAE 1

SUMMARY OF ARGUMENT 2

ARGUMENT 3

I. OpenAI's Copying for AI Training Is Not a Transformative Use. 4

A. LLM developers use copyrighted textual works for their expressive content and for the same purpose as the original works. 4

B. LLMs generate substitutes for copyrighted works and their derivatives. 5

C. No compelling justification exists for using expressive content to generate expressive content. 8

II. The Defendants' Unlicensed Uses for AI Training Substantially Harm the Markets for Copyrighted Works. 10

A. LLMs that produce substitutes for copyrighted works harm the markets for the works and their derivatives. 10

B. Unlicensed copying for training harms the expanding market for AI licensing. 13

C. AI training using works obtained from pirate sites harms existing markets and facilitates further copyright infringement. 17

CONCLUSION 18

CERTIFICATE OF WORD COUNT COMPLIANCE 20

TABLE OF AUTHORITIES

Page(s)

Constitutional Provisions

U.S. Const. art. I, § 8, cl. 8 3

Cases

A.V. v. iParadigms, LLC,
562 F.3d 630 (4th Cir. 2009) 8

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

Anderson v. Stallone,
No. 87-0592, 1989 U.S. Dist. LEXIS 11109 (C.D. Cal. Apr. 25, 1989) 7

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith,
598 U.S. 508 (2023) 3, 4, 5, 7, 8, 9, 10

Apress Media, LLC v. Anna's Archive,
No. 1:26-cv-01850, 2026 U.S. Dist. LEXIS 119530 (May 19, 2026) 18

Apress Media, LLC v. WeLib,
No. 1:26-cv-05064 (S.D.N.Y. filed June 16, 2026) 18

Authors Guild v. Google, Inc.,
804 F.3d 202 (2d Cir. 2015) 8, 11

Authors Guild, Inc. v. HathiTrust,
755 F.3d 87 (2d Cir. 2014) 8

Bartz v. Anthropic PBC,
787 F. Supp. 3d 1007 (N.D. Cal. 2025) 7, 16

Bill Graham Archives v. Dorling Kindersley Ltd.,
448 F.3d 605 (2d Cir. 2006) 2

BMG Music v. Gonzalez,
430 F.3d 888 (7th Cir. 2005) 14

Campbell v. Acuff-Rose Music, Inc.,
510 U.S. 569 (1994) 4, 8, 9, 10, 14, 16

Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.,
150 F.3d 132 (2d. Cir. 1998) 12

Dr. Seuss Enters., L.P. v. ComicMix LLC,
983 F.3d 443 (9th Cir. 2020) 16

Elsevier Inc. v. Sci-Hub,
No. 15-cv-4282, 2017 U.S. Dist. LEXIS 147462 (S.D.N.Y. June 21, 2017) 18

Fox News Network, LLC v. TVEyes, Inc.,
883 F.3d 169 (2d Cir. 2018) 14

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) 10

Harper & Row, Publishers, Inc. v. Nation Enters.,
471 U.S. 539 (1985) 10, 14

Int'l News Serv. v. Associated Press,
248 U.S. 215 (1918) 4

Kadrey v. Meta Platforms, Inc.,
788 F. Supp. 3d 1026 (N.D. Cal. 2025) 7, 12, 16, 18

Kelly v. Arriba Soft Corp.,
336 F.3d 811 (9th Cir. 2003) 8

Kirtsaeng v. John Wiley & Sons, Inc.,
579 U.S. 197 (2016) 16

Mazer v. Stein,
347 U.S. 201 (1954) 3

Monge v. Maya Mags., Inc.,
688 F.3d 1164 (9th Cir. 2012) 17

New Era Publ'ns Int'l, ApS v. Carol Publ'g Grp.,
904 F.2d 152 (2d Cir. 1990) 2

Penguin Random House LLC v. Colting,
270 F. Supp. 3d 736 (S.D.N.Y. 2017) 7

Perfect 10, Inc. v. Amazon.com, Inc.,
508 F.3d 1146 (9th Cir. 2007) 8

Romanova v. Amilus Inc.,
138 F. 4th 104 (2d Cir. 2025) 5, 9

Salinger v. Colting,
641 F. Supp. 2d 250 (S.D.N.Y. 2009),
vacated on other grounds, 607 F.3d 68 (2d Cir. 2010) 11

SunTrust Bank v. Houghton Mifflin Co.,
268 F.3d 1257 (11th Cir. 2001) 2

TCA TV Corp. v. McCollum,
839 F.3d 168 (2d Cir. 2016) 9

Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel., Inc.,
No. 25-2153, 2026 U.S. App. LEXIS 29958 (Sep. 29, 2026) 5, 10, 11, 16

Twin Peaks Prods. v. Publ'ns Int'l, Ltd.,
996 F.2d 1366 (2d Cir. 1993) 7, 12

Wainwright Secs. v. Wall St. Transcript Corp.,
558 F.2d 91 (2d Cir. 1977) 4, 12

Statutes

17 U.S.C. § 101 7

17 U.S.C. § 107 ..... 3, 9, 11

17 U.S.C. § 501 11

Other Authorities

Jasmine Abdel-khalik, *Visual Appropriation Art, Transformativeness, and Fungibility*,
48 AIPLA Q. J. 171 (2020) 9

A. Feder Cooper & James Grimmelmann, *The Files Are in the Computer: On Copyright, Memorization, and Generative AI*, 100 Chi.-Kent L. Rev. 141 (2025) 6

Daniel Gervais et al., *The Heart of the Matter: Copyright, AI Training, and LLMs*,
71 J. Copyright Soc'y 482 (2024) 4

H. Comm. on the Judiciary, 89th Cong., Supplementary Report of the Register of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill (Comm. Print 1965) 4

Melville B. Nimmer & David Nimmer, *Nimmer on Copyright* (2026) 16

U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (May 2025) (pre-publication) 12

STATEMENT OF INTERESTS OF AMICI CURIAE

The Association of American Publishers, Inc. ("AAP") is a not-for-profit organization that represents approximately 120 book, journal, and education publishers in the United States on matters of law and policy, advocating for outcomes that incentivize the publication of creative expression across a variety of formats and business models. AAP's members invest substantial time, expertise, and resources in the creation and distribution of works, including literature, scholarship, professional content, and scientific journals.

A pioneer in voluntary collective licensing, the Copyright Clearance Center ("CCC") has been dedicated to advancing copyright, accelerating knowledge, and powering innovation since 1978. CCC supports a thriving knowledge economy as a trusted intermediary, providing licensing solutions that make copyright work, including collective licensing solutions for the use of copyrighted materials with AI systems. CCC also offers a portfolio of innovative and complementary software solutions, as well as high-quality content, data, and information services.

The Independent Book Publishers Association ("IBPA") is a not-for-profit trade association founded in 1983 that represents more than 4,000 independent publishers, university and nonprofit presses, and author-publishers across the United States. IBPA's members are predominantly small and mid-sized businesses whose catalogs are disproportionately vulnerable to market dilution from AI-generated works. Lacking the scale to negotiate bilateral AI licenses, they depend on collective and aggregated licensing to participate in the developing AI licensing market.

The International Association of Scientific, Technical & Medical Publishers ("STM") is the leading global trade association for academic and professional publishers. STM members represent learned societies, university presses, and private companies of all sizes in the academic and professional publishing industry. Its members collectively publish over 60 percent of all English language journal articles and tens of thousands of monographs and reference works.

AAP, CCC, IBPA, and STM have an interest in preserving the incentives in creating and distributing works through publishing and applying the fair use doctrine as a reasonable limitation on exclusive rights, consistent with the fundamental objectives of copyright.[1]

SUMMARY OF ARGUMENT

Books and written works are foundational to the country's culture, democracy, and education. American publishers have long been catalysts for creativity and the distribution of knowledge by enabling authors to realize economic value in their works, encouraging the continued creation of works, and making distribution decisions based on the evolving marketplace. Publishers' investments in books, scholarly and professional journals, educational materials, and other textual work are essential to innovation and a well-informed and enriched public.

The Copyright Act reflects Congress's considered determination that exclusive rights, market-based licensing, and voluntary exchange encourages the creation and dissemination of books and other works. This framework remains the same no matter how much technological innovation changes the way copyrighted works are created, distributed, or used. Using copyrighted works to train generative AI models does not warrant special treatment.

Fair use is a critical exception to exclusive rights that complements the goals of copyright. Publishers themselves regularly rely on fair use for uses such as reproducing copyrighted images as historical artifacts in documentary works, *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605 (2d Cir. 2006), reproducing excerpts for purposes of criticism and biography, *New Era Publ'ns Int'l, ApS v. Carol Publ'g Grp.*, 904 F.2d 152 (2d Cir. 1990), and creating critical re-imaginings of fictional works through parody, *SunTrust Bank v. Houghton Mifflin Co.*, 268 F.3d 1257 (11th Cir. 2001). These types of transformative uses of copyrighted works benefit the public by generating new insights and information about the works without prejudicing rightsholders.

But uses that exploit written expression to generate outputs that convey similar expression or otherwise serve as substitutes are neither transformative nor fair use. Substitution is

[1] No party or its counsel authored this brief in whole or in part or contributed money toward the preparation or submission of this brief.

"copyright's bête noire." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 528 (2023). Where large language models ("LLMs"), such as OpenAI's, both share the same purpose as the copyrighted works on which they are trained and generate outputs that usurp markets that rightfully belong to authors and publishers, the risk of substitution is high, and fair use should not apply to LLM training.

Moreover, the goals of copyright are undermined when fair use is applied to authorize the uncompensated appropriation of the creative labor of thousands, if not millions, of copyright owners where licensing is practical, beneficial, and commercially reasonable. Overly expansive interpretations of fair use displace the market mechanisms Congress established, replacing negotiated exchange with uncompensated use and transferring value from the authors and publishers that invest in them to industries that neither incurred the costs nor assumed the risks of producing the underlying works. Fair use is not a roving license for courts to revise the framework Congress implemented whenever new technologies emerge and markets are still developing. In view of the multiple market harms from unauthorized uses for LLM training, courts should exercise caution to avoid foreclosing the development of voluntary licensing markets and unsettling the balance Congress established.

ARGUMENT

The Constitution empowers Congress to enact a statutory scheme to encourage and reward individual efforts in the creation and dissemination of creative works through the mechanism of exclusive rights. *See* U.S. Const. art. I, § 8, cl. 8; *Mazer v. Stein*, 347 U.S. 201, 219 (1954) (explaining the "economic philosophy behind" the Copyright Clause is "the conviction that encouragement of individual effort by personal gain is the best way to advance public welfare through the talents of authors"). The fair use doctrine under 17 U.S.C. § 107 reflects a careful balance by Congress to recognize limited circumstances permitting unauthorized uses of copyrighted works.

The technological innovation of AI is no talisman for fair use. The fair use doctrine is intended to adapt—not yield—to changes in technology. Congress expected copyright protection

to encompass unforeseen technologies, and drafted the Copyright Act with broad, technologically neutral rights and specific exceptions so that copyright owners could presumptively retain their exclusive rights even over new technological uses. *See* H. Comm. on the Judiciary, 89th Cong., Supplementary Report of the Register of Copyrights on the General Revision of the U.S. Copyright Law: 1965 Revision Bill 14 (Comm. Print 1965).

LLMs like OpenAI's exploit copyrighted works for their expressive value for the same purpose, ultimately generating substitutes for the works used for training. Fair use should not sanction AI systems with an inherent "problem of substitution." *Warhol*, 598 U.S. at 528.

I. OpenAI's Copying for AI Training Is Not a Transformative Use.

The "central" question under the first fair use factor, "the purpose and character of the use," is whether the new use "served a purpose distinct from the original, or instead superseded its objects." *Warhol*, 598 U.S. at 542 (quoting *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994)). Transformativeness requires more than mere innovation and must be justified. *See id.* at 532, 543 n.18. A generative AI model that sources expressive content to generate outputs containing expressive content is a tool with little transformative purpose.

A. LLM developers use copyrighted textual works for their expressive content and for the same purpose as the original works.

OpenAI and other LLM developers highly value books and other textual works for their expressive content. Protected expression encompasses "the particular form or collocation of words in which the writer has communicated it," *Wainwright Secs. v. Wall St. Transcript Corp.*, 558 F.2d 91, 95 (2d Cir. 1977) (quoting *Int'l News Serv. v. Associated Press*, 248 U.S. 215, 234 (1918))—precisely the "linguistic structure" OpenAI's LLM training seeks to capture at scale from copyrighted works, OpenAI's Mem. 20 (Dkt. 1983, 25-md-3143). As AI researchers have noted, "books are invaluable for long-range context modeling research and coherent storytelling." Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* 4 (Dec. 31, 2020), https://arxiv.org/pdf/2101.00027; *see also* Daniel Gervais et al., *The Heart of the Matter: Copyright, AI Training, and LLMs*, 71 J. Copyright Soc'y 482, 505 (2024) ("[T]he reason for

ingesting copyrighted works into the encoder layer of an AI model is directly and unequivocally due to its expressive and aesthetic value."). Published works are especially valued for AI training because they have benefited from substantial human investment in research, editing and peer review, and validation.

A use that "shares the purpose" of an original work is more likely to be a "substantial substitute" for that work or its derivative, "which undermines the goal of copyright." *Warhol*, 598 U.S. at 531-32 (citation modified). Creators use their expression in books and other copyrighted works to connect with readers to inform, educate, and entertain. Likewise, OpenAI uses the expression in copyrighted works to create a platform to connect with users to inform, educate, and entertain. As the Third Circuit recently held, using copyrighted works "to train an AI program" for the platform's benefit does not deter a finding that the defendant "shares the same ultimate purpose" as the creator. *Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel., Inc.*, No. 25-2153, 2026 U.S. App. LEXIS 29958, at *17-18 (Sep. 29, 2026). And where generating "short stories, long form journalism, scientific articles, textbooks, or novels" was an "important future direction" for OpenAI, the purpose is plainly the same. Class Pls.' Statement of Undisputed Material Facts ¶ 708 ("Pls.' SUF") (Dkt. 1987, 25-md-3143) (quoting Nelson Decl. Ex. 283).

B. LLMs generate substitutes for copyrighted works and their derivatives.

A use is transformative if it does not serve as a substitute for the original. *See Romanova v. Amilus Inc.*, 138 F. 4th 104, 110-11 (2d Cir. 2025). Consistent with their purpose and training, LLMs can readily generate substitutes for the copyrighted works on which they are trained, resulting in non-transformative uses. LLM outputs used in environments that are not "distinct and different" from those of the original works supersede their objects. *Warhol*, 598 U.S. at 535-36 (quoting *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 31 (2021)).

The substitutes generated by LLMs can take the form of verbatim and near-verbatim copies, summaries and alternative versions of written works, knock-offs that copy expression and creative choices from original works, and derivative works exclusively reserved for rightsholders.

LLMs' ability to generate such outputs or otherwise exploit the expressive content of the original works demonstrates that LLMs are not highly transformative.

Studies have shown that LLMs—including OpenAI's GPT models—can memorize and reproduce in outputs substantial portions of books that were used for training. Researchers recently found that "finetuning bypasses [the] protections" implemented by AI companies to prevent regurgitation of copyrighted works and that training models such as GPT-4o in certain ways "causes them to reproduce up to 85-90% of held-out copyrighted books, with single verbatim spans exceeding 460 words." Xinyue Liu et al., *Alignment Whack-a-Mole: Finetuning Activates Verbatim Recall of Copyrighted Books in Large Language Models*, 2026 Conf. on Language Modeling 1, https://arxiv.org/pdf/2603.20957. Others have found that, despite the use of safeguards, the extraction of texts used for training is still a risk with LLMs. Ahmed Ahmed et al., *Extracting Books from Production Language Models* 20 (Jan. 6, 2026), https://arxiv.org/pdf/2601.02671 ("LLMs memorize portions of their training data, these memorized data are encoded in the model's weights, and, as we show here, it can be feasible to extract large quantities of in-copyright training data from production LLMs." (citations omitted)); *see also* A. Feder Cooper & James Grimmelmann, *The Files Are in the Computer: On Copyright, Memorization, and Generative AI*, 100 Chi.-Kent L. Rev. 141, 210 (2025) (concluding that a generative AI model "stores near-exact copies of portions of works it was trained on," regardless of its capability to output the works).

OpenAI's own documents confirm that the nature of its LLMs is to regurgitate: "[M]odels want to regurgitate, so we're inherently fighting against the training objective." Pls.' SUF ¶ 559 (quoting Nelson Decl. Ex. 298); *see also id.* ¶ 563 (noting that a "major problem" with GPT-3.5 is that it "can easily reproduce training data verbatim" (quoting Nelson Decl. Ex. 292)). Relying on safeguards to resist regurgitation has become questionable with recent reports from OpenAI and other developers about models diverging from their design and disabling safeguards. *See Our Framework for Reporting Model Misalignment*, OpenAI (Sep. 16, 2026), https://openai.com/index/model-misalignment-reporting-framework/; *Investigating Three Real-*

World Incidents in Our Cybersecurity Evaluations, Anthropic (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals. If alarms are sounding off for cyber and national security, what does this say about OpenAI's controls for copyright?

Regurgitation aside, the defendants do not dispute that ChatGPT can provide short summaries and other outputs adapted from copyrighted works. *See* OpenAI's Mem. 29-30. For instance, ChatGPT has generated sequels that are substitutes for George R. R. Martin's much-anticipated book, *The Winds of Winter*. Pls.' SUF ¶¶ 736-739 (citing Shan Decl. Ex. D). ChatGPT's substitutive character was enabled not only by training on a pirated copy of Martin's *A Song of Ice and Fire*, *see id.* ¶ 250 (citing Nelson Decl. Exs. 828, 830), but also by pursuing a "research mission" to develop an LLM that can "write the last two books of . . . *A Song of Ice And Fire*," *id.* ¶ 727 (citing Nelson Decl. Exs. 716, 746).

That LLMs can be used and are designed to produce outputs that may qualify as derivative works implicates rights reserved to copyright owners and requires more to be transformative. *See Warhol*, 598 U.S. at 529. Unauthorized derivative works can include detailed summaries, *see Twin Peaks Prods. v. Publ'ns Int'l, Ltd.*, 996 F.2d 1366, 1373 (2d Cir. 1993) (guidebook recounting episodes of the "Twin Peaks" teleplay with "[e]very intricate plot twist and element of character development"); sequels, *see Anderson v. Stallone*, No. 87-0592, 1989 U.S. Dist. LEXIS 11109, at *24-25 (C.D. Cal. Apr. 25, 1989) (*Rocky IV* film containing the distinctive Rocky character with the same name, relationships, and experiences from prior *Rocky* films); and other adaptations, *see Penguin Random House LLC v. Colting*, 270 F. Supp. 3d 736, 748-49 (S.D.N.Y. 2017) (illustrated children's versions of adult novels "basically retell[ing] the story" of the original books); *see also* 17 U.S.C. § 101 (providing examples in the "derivative work" definition).

Kadrey v. Meta Platforms, Inc., 788 F. Supp. 3d 1026 (N.D. Cal. 2025) and *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025), are not controlling. In deciding transformativeness, those cases did not fully consider LLMs' ability to generate substitutes of copyrighted works, such as derivatives, or the same facts present here.

Other cases often cited by AI developers and their supporters also do not apply to LLM training. Fair use has been found where copying enabled technological tools that analyze and provide new information about the individual copyrighted works or the corpus as a whole. *See, e.g.*, *Authors Guild v. Google, Inc.* ("*Google Books*"), 804 F.3d 202, 217 (2d Cir. 2015) ("allow[ing] readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods"); *A.V. v. iParadigms, LLC*, 562 F.3d 630, 638 (4th Cir. 2009) (detecting plagiarism of students' written works). Fair use also has been found for search and retrieval tools that direct users to relevant works or portions of those works. *See, e.g.*, *Google Books*, 804 F.3d at 217-25 (displaying tiny book snippets with terms of interest); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97-101 (2d Cir. 2014) (listing books with terms of interest); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1155, 1164-68 (9th Cir. 2007) (displaying thumbnail, lower-resolution images as links to works); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818-22 (9th Cir. 2003) (same). But neither of those sets of cases involved copying that exploited expressive content for the purpose of generating expressive outputs. While LLM training may appear technically similar to the tools in those cases, LLMs have a fundamentally different relationship with copyrighted works as to what they extract and output and therefore can provide substitutes for the expressions in those works.

C. No compelling justification exists for using expressive content to generate expressive content.

A compelling justification is "particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it." *Warhol*, 598 U.S. at 532. A justification requirement harkens back to traditional fair use applications "to works whose very subject is the original composition and so necessitates some borrowing from it." *Campbell*, 510 U.S. at 597 (Kennedy, J., concurring). For example, "commentary or criticism on another's work is among the best recognized justifications for copying" in part because such a use "frequently requires quoting the copyrighted work to be

effective." *TCA TV Corp. v. McCollum*, 839 F.3d 168, 180 n.10 (2d Cir. 2016) (citation modified). The Copyright Act identifies other paradigmatic examples of fair use that, similar to commentary and criticism, provide insight or information about the original work. *See* 17 U.S.C. § 107 (referring to news reporting, teaching, scholarship, and research). The Supreme Court explained in *Campbell* that, if a new work "has no critical bearing on the substance or style of the original composition, which the alleged infringer merely uses to get attention or to avoid the drudgery in working up something fresh, the claim to fairness in borrowing from another's work diminishes accordingly (if it does not vanish)." 510 U.S. at 580. *Warhol* went further to state that a "particularly compelling justification is needed" where the defendant's use is "so similar" to the "typical use" of the original work. 598 U.S. at 547.

There is no apparent, let alone compelling, justification for AI training on expressive content to generate expressive content. As proponents of an affirmative defense "bear[ing] the burden of justifying [their] taking" of copyrighted works, *Romanova*, 138 F.4th at 110, the defendants have not shown why unauthorized copying of copyrighted works is necessary. Certainly, no justification has been given for using any particular copyrighted work in OpenAI's training sets. OpenAI contends that LLMs learn by analyzing vast amounts of texts and observing "patterns corresponding to linguistic rules (*e.g.*, subject-verb agreement) or common facts (*e.g.*, the name of our second President)." OpenAI's Mem. 9-10. But other texts would have been interchangeable for these purposes; OpenAI did not need to use the copyrighted works at issue. *See* Jasmine Abdel-khalik, *Visual Appropriation Art, Transformativeness, and Fungibility*, 48 AIPLA Q. J. 171, 222 (2020) ("The more the characteristics of the original work are fungible for the purposes of the secondary user, the less likely the secondary user should have a claim to fair use."). Indeed, the defendants disavow any need for the specific works used for training. *See, e.g.*, Pls.' SUF ¶ 497 (OpenAI expert stating that "any individual source . . . is entirely fungible" (quoting Nelson Decl. Ex. 751)).

Further, researchers question whether bigger is better for LLM development: while more weights may work in the short term, frontier AI companies that "place disproportionate emphasis

on scaling laws are likely under-investing in other directions of innovation which will unlock future gains." Sara Hooker, *On the Slow Death of Scaling* 9 (Dec. 6, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5877662. Improved model performance can be achieved through other means, with less data of higher quality, algorithmic improvements, or a new architectural design. *See id.* at 6-7. Using copyrighted works merely to avoid the costs, time, or "drudgery" of working on alternatives weighs against fair use. *Campbell*, 510 U.S. at 580; *see also Thomson Reuters*, 2026 U.S. App. LEXIS 29958, at *21 ("[E]ase is not a justification for copying.").

II. The Defendants' Unlicensed Uses for AI Training Substantially Harm the Markets for Copyrighted Works.

The fourth fair use factor, "the effect of the use upon the potential market for or value of the copyrighted work," is "undoubtedly the single most important element of fair use." *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 566 (1985); *accord Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 189 (2d Cir. 2024). This is so regardless of how transformative LLMs are considered to be under the first factor. *See Warhol*, 598 U.S. at 536 n.12.

Courts must account for harms to the original work and to the market for derivative works, along with market effects from unrestricted and widespread unauthorized uses similar to the defendants'. *Campbell*, 510 U.S. at 590. The markets harmed from unauthorized AI training include the markets for actual and potential substitutes, including derivative works, and for AI licensing.

A. LLMs that produce substitutes for copyrighted works harm the markets for the works and their derivatives.

Unauthorized copying for AI training harms the potential market for and value of the copyrighted original where the resulting LLM enables widespread access to outputs that reproduce significant protected expression or otherwise provide a substantial substitute for the work or its derivative. As discussed above, unlike a system limited to search, indexing, or constrained

snippets, OpenAI's generative systems can produce expressive outputs satisfying user demand and do not favor a fair use finding.

In particular, LLMs far exceed the "boundaries of fair use" that were tested in *Google Books*. 804 F.3d at 206. The Second Circuit's findings as to Google's search and snippet functions were premised on providing those functions "without providing the public with a substantial substitute for matter protected by the Plaintiffs' copyright interests in the original works or derivatives of them." *Id.* at 207. The court noted that, even if the copying was for a transformative purpose, "such copying might nonetheless harm the value of the copyrighted original if done in a manner that results in widespread revelation of sufficiently significant portions of the original as to make available a significantly competing substitute." *Id.* at 223. In contrast to Google's snippet feature, which minimized the aggregate amount of viewable copyrighted text, OpenAI's safeguards can be bypassed, as discussed above, to extract substantial amounts that serve as effective, free substitutes for the original work.

Because the standards for a fair use affirmative defense, 17 U.S.C. § 107, and an infringement claim, *id.* § 501, are different, LLM outputs need not be substantially similar to the original works to result in market harm. Indeed, substantial similarity was neither mentioned nor required by the Third Circuit in finding that "unrestricted and widespread copying" of the plaintiffs' headnotes "to develop substitute legal-research tools" would harm the plaintiff in the market. *Thomson Reuters*, 2026 U.S. App. LEXIS 29958, at *25-26 (citation modified).

Unquestionably, LLMs that output unauthorized sequels, translations, detailed summaries, and other derivative works harm the rightsholder's derivative market. For example, generating a novel that "continues the story" of an original work and "similar widespread works could substantially harm the market for [the original work's] sequel or other derivative works, whether through confusion as to which is the true sequel or companion to [the original work], or simply because of reduced novelty or press coverage." *Salinger v. Colting*, 641 F. Supp. 2d 250, 267-68 (S.D.N.Y. 2009), *vacated on other grounds*, 607 F.3d 68 (2d Cir. 2010). While the defendants doubt the plausibility of someone reading a summary in lieu of a book, *see* OpenAI's Mem. 30,

the Second Circuit has long recognized that a summary may be "an adequate substitute," foregoing the need to consume the original work. *Twin Peaks Prods.*, 996 F.2d at 1377; *see also Wainwright Secs.*, 558 F.2d at 93, 96 (finding the defendant's abstracts had the "obvious intent, if not the effect, of fulfilling the demand" for the plaintiff's lengthy financial reports). Similarly, a book containing questions and answers about the events and characters of a popular work may well "satiate" fans' passion for the work and substitute for a derivative market. *Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 143, 145-46 (2d. Cir. 1998). Services that provide summaries as licensed derivatives have long existed, but they are quickly being supplanted by LLMs that handle both the reading and analysis of books. *See* Naomi S. Baron, *AI Is Making Reading Books Feel Obsolete – and Students Have a Lot to Lose*, The Conversation (Aug. 21, 2025), https://theconversation.com/ai-is-making-reading-books-feel-obsolete-and-students-have-a-lot-to-lose-262680. A user-prompted AI tool that can generate outputs derived from expression in copyrighted works, such as summaries, translations, answers to questions, or newly-produced text, thus can satisfy user demand for the works themselves.

Another harmful effect on the potential market for or value of copyrighted works are indirect substitutes generated by LLMs. Similar to derivative market harms, consumer confusion as to authorship or reader overload can result in market displacement from LLMs outputs that do not reproduce any part of the works used in training. "[AI] involves a technology that can generate literally millions of secondary works, with a miniscule fraction of the time and creativity used to create the original works it was trained on. No other use . . . has anything near the potential to flood the market with competing works the way that LLM training does." *Kadrey*, 788 F. Supp. 3d at 1055; *see also* U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 65 (May 2025) (pre-publication).

The harmful market effects of LLM outputs are significant. The number of e-books published per month has tripled since the release of ChatGPT. *See* Imke Reimers & Joel Waldfogel, *AI and the Quantity and Usage of Creative Products: Have LLMs Boosted Creation of Valuable Books?* 2, 4 (Nat'l Bureau of Econ. Rsch., Working Paper No. 34777, 2026),

https://www.nber.org/system/files/working_papers/w34777/w34777.pdf. AI-generated books are permeating the markets, *see, e.g.*, Michael Fraiman, *63% of All Religious Books on Amazon Are Likely AI-Written*, Originality.ai (Aug. 19, 2026), https://originality.ai/blog/religious-books-on-amazon-study, and reports of blatant AI-generated rip-offs abound at major booksellers, *see, e.g.*, CT Jones, *Amazon Is the World's Biggest Online Book Marketplace. It's Filled with AI Knockoffs*, Rolling Stone (Oct. 27, 2025), https://www.rollingstone.com/culture/culture-features/amazon-ai-book-knockoffs-1235450690/; Devesh Beri, *Apple Books Is Getting Flooded With AI-Generated Plagiarism*, ExtremeTech (June 29, 2026), https://www.extremetech.com/internet/apple-books-is-getting-flooded-with-ai-generated-plagiarism. Where writing a single book typically would take years, a single individual used ChatGPT to produce over 400 books in the span of 17 months to later sell on Amazon. Kashmir Hill, *I Got Slopped*, The N.Y. Times (July 16, 2026), https://www.nytimes.com/2026/07/16/technology/ai-slop-books-biography-amazon.html. Consumers are confused and overwhelmed, and publishers risk losing revenues and market visibility to the flood of AI-generated books and other texts built on the backs of their own works.

The defendants themselves have acknowledged these market dilution harms. An OpenAI researcher forewarned that it would become "impossible to tell" whether material was AI-generated given his goal to "make the most artisanal slop." Pls.' SUF ¶ 733 (quoting Nelson Decl. Ex. 722). A Microsoft employee also concluded that the economic harm from LLMs included "authors . . . who were no longer able to sell their books because there are so many substitutive AI-generated books out there." *Id.* ¶¶ 767-68 (quoting Nelson Decl. Ex. 26).

B. Unlicensed copying for training harms the expanding market for AI licensing.

Exclusive rights under the Copyright Act anticipate that mutually beneficial agreements for most uses of copyrighted works will be reached through the marketplace. Licensing copyrighted works is a well-established mechanism benefiting publishers that heavily invest in human ingenuity, research, judgment, and expertise in the creation of those works; licensing for AI uses is a natural extension of that established framework. Fair use should not be used to

displace a "fully functioning market that encourages the creation and dissemination" of creative works, *Harper*, 471 U.S. at 566 n.9, or to allow AI companies that refuse to participate in the marketplace to thwart the development of the growing AI licensing market, *see BMG Music v. Gonzalez*, 430 F.3d 888, 891 (7th Cir. 2005) ("[C]opiers . . . cannot ask courts (and juries) to second-guess the market and call wholesale copying 'fair use.'").

Campbell clarified the scope of cognizable market harm: whereas disparagement is unremediable, displacement or usurpation of demand for the original work is remediable. *See* 510 U.S. at 591-92. So is harm to a market that "creators of original works would in general develop or license others to develop." *Id.* at 592; *see also Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 929-30 (2d Cir. 1994) (considering markets that are "traditional, reasonable, or likely to be developed"); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 180 (2d Cir. 2018) (finding displacement of licensing revenues in "plausibly exploitable market" for access to copyrighted content).

Here, a market for licensing copyrighted works for AI exists and is expanding to address growing demands and developing infrastructure. The list of publicly reported AI licensing arrangements is extensive, covering works of all types and AI developers of all sizes. *See AI Licensing by Copyright Owners*, Copyright All., https://copyrightalliance.org/artificial-intelligence-copyright/licensing/copyright-owners/ (last visited Oct. 27, 2026). Publishers of books, journals, and educational materials are active in this market, with licenses spanning a

spectrum of uses from LLM training[2] to retrieval-augmented generation ("RAG")[3] to AI search[4] and product integration[5]. Other licensing frameworks are being offered in addition to direct licenses. For example, CCC offers a suite of collective licensing solutions that include AI rights, including an enterprise-wide license offering rights to millions of text-based works for internal uses, a campus-wide license for higher education uses, and a license specifically for training external AI systems. *See, e.g.*, *Annual Copyright License*, Copyright Clearance Ctr., https://www.copyright.com/solutions-annual-copyright-license/ (last visited Oct. 27, 2026); *AI*

---

[2] *See, e.g.*, Press Release, Wiley, *Wiley Partners with Pi School to Enhance the European Space Agency's AI-Powered Earth Virtual Expert (EVE) Project* (Mar. 12, 2025), https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-Partners-with-Pi-School-to-Enhance-the-European-Space-Agencys-AI-Powered-Earth-Virtual-Expert-EVE-Project/default.aspx; Hannah Miller & Dina Bass, *Microsoft Signs AI-Learning Deal with News Corp.'s HarperCollins*, Bloomberg (Nov. 19, 2024), https://www.bloomberg.com/news/articles/2024-11-19/microsoft-signs-ai-learning-deal-with-news-corp-s-harpercollins.

[3] *See, e.g.*, Press Release, OpenEvidence, *Springer Nature and OpenEvidence Announce Agreement to Maximise Exposure of Trusted Findings on OpenEvidence Platform* (Aug. 5, 2026), https://www.openevidence.com/announcements/springer-nature-and-openevidence-announce-agreement-to-maximise-exposure-of-trusted-findings-on-openevidence-platform; Press Release, OpenEvidence, *OpenEvidence and NEJM Group, Publisher of the New England Journal of Medicine, Sign Content Agreement* (Feb. 19, 2025), https://www.openevidence.com/announcements/openevidence-and-nejm; Press Release, Informa, *Informa PLC* 3 (May 8, 2024), https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc---market-update.pdf.

[4] *See, e.g.*, *De Gruyter Brill Partners with Consensus to Enhance Research Discovery*, De Gruyter Conversations (June 8, 2026), https://blog.degruyter.com/de-gruyter-brill-partners-with-consensus-to-enhance-research-discovery/; Press Release, Wiley, *Wiley and Perplexity Announce New AI Search Partnership* (May 8, 2025), https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-and-Perplexity-Announce-New-AI-Search-Partnership/default.aspx.

[5] *See, e.g.*, Steven Johnson & Will Houghteling, *Expert Intelligence: A New Way for You to Engage with Trusted Content*, News from Google (Aug. 27, 2026), https://blog.google/innovation-and-ai/products/gemini-notebook/expert-intelligence-leading-sources/ (launching ebook engagement tool with licensed content from major publishers, including Bloomsbury, De Gruyter Brill, Johns Hopkins University Press, Macmillan Publishers, and Penguin Random House); Kevin Davies, *Sage Inks Licensing Deal with Causaly's AI Platform Focusing on Drug Discovery Research*, Genetic Eng'g & Biotechnology News (July 29, 2026), https://www.genengnews.com/industry-news/sage-inks-publishing-deal-with-causalys-ai-platform-focusing-on-drug-discovery-research/.

Systems Training License, Copyright Clearance Ctr., https://www.copyright.com/solutions-ai-systems-training-license/ (last visited Oct. 27, 2026). Publishers and AI content management platforms have formed partnerships to license catalogs of textual works. *See, e.g.*, Ed Nawotka, *Princeton University Press Partners with Cashmere on AI Licensing*, Publishers Weekly (Aug. 3, 2026), https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/100971-princeton-university-press-partners-with-cashmere-on-ai-licensing.html. The AI training market for licensing for all content types was valued at $5.2 billion in 2025, with text holding the largest share of the market, and is projected to rapidly reach $55.8 billion by 2035. *See AI Training Dataset Market*, Vantage Mkt. Rsch. (May 2026), https://www.vantagemarketresearch.com/ai-training-dataset-market.

The AI licensing market thus is far from the type of "theoretical market" where the concern of "circularity" arises. *See* 4 Melville B. Nimmer & David Nimmer, *Nimmer on Copyright* § 13F.08[B] (2026). "This is not a case where the copyist's work fills a market that the copyright owner will likely avoid," *Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 460 (9th Cir. 2020), or where there is evidence of market failure. The district courts in *Kadrey* and *Bartz* failed to properly consider the AI licensing market as required by *Campbell*. *See* 788 F. Supp. 3d at 1052; 787 F. Supp. 3d at 1032. Applying *Campbell*, the Third Circuit found the licensing market for AI training as "rapidly developing" and harmed by the defendant's unauthorized copying. *Thomson Reuters*, 2026 U.S. App. LEXIS 29958, at *26-27.

Recognizing an AI licensing market further has the benefit of serving the subsidiary aims of copyright: "Copyright law ultimately serves the purpose of enriching the general public through access to creative works. The statute achieves that end by striking a balance between two subsidiary aims: encouraging and rewarding authors' creations while also enabling others to build on that work." *Kirtsaeng v. John Wiley & Sons, Inc.*, 579 U.S. 197, 204 (2016) (citation modified).

Licensing promotes the continued creation and dissemination of new and previously-published works by compensating authors and publishers and giving them a say in whether and how their works are exploited. The U.S. publishing industry supports an extensive network of

American businesses and thousands of jobs, with books and course materials generating over $33 billion in 2025. *See* Press Release, Ass'n of Am. Publishers, *AAP StatShot Annual Report: Publishing Revenues Totaled $33.4 Billion for Calendar Year 2025* (Aug. 10, 2026), https://publishers.org/news/aap-statshot-annual-report-publishing-revenues-totaled-33-4-billion-for-calendar-year-2025/. The publishing industry is an integral part of broader U.S. copyright industries, which in 2023 collectively contributed more than $2 trillion to the U.S. economy. Jéssica Dutra & Robert Stoner, Int'l Intell. Prop. All., *Copyright Industries in the U.S. Economy: The 2024 Report* 8 (2025), https://www.iipa.org/files/uploads/2025/02/IIPA-Copyright-Industries-in-the-U.S.-Economy-Report-2024_ONLINE_FINAL.pdf. The revenues publishers earn through the licensing and distribution of their works are used to compensate authors and invest in the next generation of books, professional content, and learning solutions.

Licensing also provides AI developers authorized access to human-created works and high-quality content. Using unauthorized content from online or questionable sources carries the risk of tainting models with low-quality, such as AI-generated, text. For AI to flourish, licensing is needed to ensure humans continue to be incentivized to create so that the well of human creation does not run dry and AI companies have reliable access to content that improves model performance. As an OpenAI researcher aptly stated: "Our AI content strategy has started a doom loop that will hurt the performance of our models and the entire web at the same time. . . . LLMs are useful because they replicate the patterns in the training content, which without ways of distributing economic value down the supply chain, necessarily threatens the economic sustainability of those who create the content." Pls.' SUF ¶ 765 (quoting Nelson Decl. Ex. 26).

C. AI training using works obtained from pirate sites harms existing markets and facilitates further copyright infringement.

"Unrestricted and widespread reproduction" similar to OpenAI's conduct not only undermines the ability of publishers to market and license authors' works, "but would also create incentives to pirate intellectual property." *Monge v. Maya Mags., Inc.*, 688 F.3d 1164, 1182 (9th

Cir. 2012); *see also Kadrey*, 788 F. Supp. 3d at 1047 (noting the relevance of effects of pirated copies "insofar as they could contribute to use of the libraries and thus infringement by others").

Pirate sites similar to LibGen have been the subject of several enforcement actions by publishers and judgments of willful copyright infringement. *See Apress Media, LLC v. Anna's Archive*, No. 1:26-cv-01850, 2026 U.S. Dist. LEXIS 119530, at *3 (May 19, 2026); *Elsevier Inc. v. Sci-Hub*, No. 15-cv-4282, 2017 U.S. Dist. LEXIS 147462, at *3 (S.D.N.Y. June 21, 2017); Motion for Default Judgment, *Apress Media, LLC v. WeLib*, No. 1:26-cv-05064 (S.D.N.Y. Oct. 12, 2026). What had been a "dying" business has not only been revived but also actively exploited and proliferated by OpenAI and other LLM developers seeking quick and easy access to training data. *See* Margaux Poueymirou & Maxwell V. Pritt, *Why AI's Use of Shadow Libraries Should Alarm Us All*, Daily J. (June 16, 2025), https://www.dailyjournal.com/articles/386130-why-ai-s-use-of-shadow-libraries-should-alarm-us-all. Anna's Archive, for example, is profiting from LLM developers willing to pay "donations" as high as $200,000 for high-speed access to its collections. Ed Newton-Rex, LinkedIn (Jan. 22, 2026), https://www.linkedin.com/posts/ed-newton-rex_a-few-months-ago-i-was-curious-to-know-how-share-7420048306082398208-XxVE/. AI companies' use of books and other copyrighted text from illegal sources normalizes large-scale theft by supporting and promoting the expansion of piracy as a business.

CONCLUSION

The Court should grant the plaintiffs', and deny the defendants', motions for summary judgment and find that OpenAI's unauthorized use of copyrighted works for AI training is not fair use.

Dated: October 29, 2026

s/ Lucy Grace D. Noyola

Lucy Grace D. Noyola (pro hac vice pending)
ASSOCIATION OF AMERICAN PUBLISHERS
1730 Pennsylvania Avenue, NW, Suite 300
Washington, DC 20006
Telephone: (202) 220-4546
lnoyola@publishers.org

Attorney for Amici Curiae
Association of American Publishers,
Copyright Clearance Center,
Independent Book Publishers Association, and
International Association of Scientific, Technical & Medical Publishers

CERTIFICATE OF WORD COUNT COMPLIANCE

In accordance with Local Civil Rule 7.1(c) and the Court's August 7, 2026 Order (Dkt. 1663, 25-md-3143), I certify that the foregoing amicus brief contains 5,998 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/ Lucy Grace D. Noyola
Lucy Grace D. Noyola