UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: All Cases	No. 25-md-3143 (SHS) (OTW)

BRIEF OF AMICI CURIAE AI COALITION FOR DATA INTEGRITY ET AL. IN SUPPORT OF PLAINTIFFS' MOTIONS FOR SUMMARY JUDGMENT AND IN OPPOSITION TO DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT

Tod Cohen (*pro hac vice* motion pending, ECF No. 2115)
LAW OFFICE OF TOD COHEN
1159 Church Street
San Francisco, CA 94114
(408) 410-8329
tcohen@gmail.com
Counsel for Amici Curiae

TABLE OF CONTENTS

Page

TABLE OF AUTHORITIES ii

INTEREST OF AMICI CURIAE 1

INTRODUCTION 1

SUMMARY OF ARGUMENT 2

ARGUMENT 4

I. NATIONAL SECURITY, PROPERLY WEIGHED UNDER FACTORS ONE AND FOUR, CUTS AGAINST UNPAID AND UNDOCUMENTED TRAINING DATA 4

A. Professionally Produced Information Is Economic Infrastructure, and a Supply Chain That No One Pays For Will Not Survive 4

B. The Government's Own Security Agencies Call for Documented Training Data 7

C. National Security Also Weighs Against Fair Use Under the First Factor 8

II. COURTS AND REGULATORS ABROAD CONDITION AI TRAINING ON LAWFUL ACCESS AND RIGHTSHOLDER CHOICE, SO A FAIR USE RULING WOULD NOT GIVE U.S. DEVELOPERS THE ADVANTAGE THE DOJ PREDICTS 9

III. THE TRAINING-DATA MARKET IS REAL, AND THE OPT-OUT DEFENSE CANNOT SUBSTITUTE FOR IT 12

A. The Market Exists, and the Defendants Participate in It 12

B. The Circularity Objection Fails Where, as Here, There Are Real Buyers, Real Sellers, and a Clearance Mechanism 13

C. Opacity Suppresses the Market, and an Opt-Out Cannot Cure It 15

D. Opacity Entrenches Incumbents; Disclosure and Licensing Promote Competition 16

IV. THE COURT SHOULD ANALYZE EACH USE SEPARATELY, ON A RECORD OF WHAT WAS USED, AND SHOULD NOT REWARD DEVELOPERS FOR KEEPING NO RECORDS 17

CONCLUSION 18

TABLE OF AUTHORITIES

Page(s)

Cases

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

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).....3, 17

Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537 (S.D.N.Y. 2013).....16, 17

Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).....14, 17

Bartz v. Anthropic PBC, 787 F. Supp. 3d 1007 (N.D. Cal. 2025).....10, 13, 14, 18

Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605 (2d Cir. 2006).....13

Cambridge Univ. Press v. Patton, 769 F.3d 1232 (11th Cir. 2014).....14

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).....3, 6, 18

Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018).....17

GEMA v. OpenAI, Landgericht München I, Az. 42 O 14139/24 (Nov. 11, 2025).....3, 9

GEMA v. Suno, Landgericht München I, Az. 42 O 763/25 (July 31, 2026).....3, 9, 10

Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021).....4, 6, 18

Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163 (2d Cir. 2024).....2, 6

Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539 (1985).....3, 17

Kadrey v. Meta Platforms, Inc., 788 F. Supp. 3d 1026 (N.D. Cal. 2025).....10, 13, 14

NXIVM Corp. v. Ross Inst., 364 F.3d 471 (2d Cir. 2004).....3, 17, 18

Pac. & S. Co. v. Duncan, 744 F.2d 1490 (11th Cir. 1984).....17

Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel. Inc., No. 25-2153 (3d Cir. Sept. 29, 2026).....passim

X.AI LLC v. Bonta, No. 26-1591 (9th Cir.).....1

Statutes and Regulations

17 U.S.C. § 106(1).....15

17 U.S.C. § 107.....3, 10

Chosakuken-hō [Copyright Act], Act No. 48 of 1970, art. 30-4 (Japan).....10

Directive 2019/790, art. 4, 2019 O.J. (L 130) 92.....10

Interim Measures for the Management of Generative Artificial Intelligence Services, art. 7(1) (China 2023).....10

Jeojakgwonbeop [Copyright Act], art. 35-5 (S. Kor.).....10

Regulation (EU) 2024/1689, art. 53 & recital 106, 2024 O.J. (L 2024/1689).....11

Other Authorities

Austl. Signals Directorate's Austl. Cyber Sec. Ctr. et al., *Artificial Intelligence and Machine Learning: Supply Chain Risks and Mitigations* (Mar. 2026).....7

KI Hub, GEMA, https://www.gema.de/en/about-gema/ki-hub (last visited Oct. 30, 2026).....11

Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 Nature 755 (2024).....6

Ministry of Culture, Sports & Tourism & Korea Copyright Comm'n (S. Kor.), *Guide on Applicability of the Fair Use Doctrine to Training of Generative AI Models* (Feb. 2026).....10

Nat'l Sec. Agency et al., *AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems* (May 22, 2025).....7

Press Release No. 11, LG München I (Nov. 11, 2025).....9

Press Release No. 16, LG München I (July 31, 2026).....10

Shayne Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons*, *in* 37 Advances in Neural Information Processing Systems 108042 (2024).....16

U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (pre-publication version May 2025).....12, 14, 16, 18

INTEREST OF AMICI CURIAE

The AI Coalition for Data Integrity ("AICDI") is a coalition of technology and content interests that supports transparency about artificial intelligence ("AI") training data and the licensing of copyrighted works used in training. AICDI has filed amicus briefs on these issues in *Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.*, No. 25-2153 (3d Cir.), and *X.AI LLC v. Bonta*, No. 26-1591 (9th Cir.). AICDI is joined by Writers Guild of America, West, Inc.; Writers Guild of America East, Inc.; National Association of Voice Actors; Microstocksolutions LLC; Digital Media Licensing Association, Inc. d/b/a Digital Media Leadership Alliance; Transparency Coalition AI; Association of Medical Illustrators; Association of Independent Music Publishers; and Music Workers Alliance, Inc. Each amicus's interest, and its corporate disclosure, is set out in the memorandum of law accompanying the motion for leave.

No party's counsel authored this brief in whole or in part, and no party or party's counsel contributed money intended to fund its preparation or submission.

INTRODUCTION

The United States, through the Department of Justice ("DOJ"), asks the Court to weigh national security in the fair use balance. Amici take no position on whether it belongs there. They write to show that, if the Court considers it, the security interest favors Plaintiffs on this record, and that the same is true of the international and market considerations the United States raises.

AICDI joins here with organizations of the writers, performers, and creators whose works are used to train AI models. Amici participate in the licensing market that Defendants say cannot work at scale. Amici's interest is an AI training-data market that works: developers know what they are building on, creators are paid for what they supply, and a record of both exists.

The Defendants built their models the opposite way. Plaintiffs' evidence indicates works taken from pirate sites, from a non-commercial license, and from search crawls that publishers allowed only for search. It shows copyright notices stripped from training copies, an announced opt-out tool abandoned, and training data traded between the Defendants themselves. The Defendants ask the Court to hold that this conduct was fair use.

That holding would reach well beyond these parties. It would weaken the market that pays for the reporting, books, and scholarship on which the models depend. It would reward developers that disclose no record of what they copied, even as the government's own security agencies urge organizations to track where training data came from. And it would put U.S. law out of step with other jurisdictions, where a court has already held that OpenAI's models reproduce protected works. As the Third Circuit recently held, incorporating AI does not give a developer "carte blanche to violate copyright law." *Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel. Inc.*, No. 25-2153, slip op. at 28 (3d Cir. Sept. 29, 2026).

SUMMARY OF ARGUMENT

I. If the Court weighs national security, it should weigh all of it. The government's own security agencies call for tracking the provenance of training data, because a corpus no one can account for cannot be checked for poisoning or manipulation. A licensed corpus comes with that record. Mass unlicensed copying leaves no comparable record. Generative AI also depends on a continuing supply of professionally produced information, and a Microsoft document in the record warns that Microsoft's own AI content strategy has started a "doom loop." The Second Circuit has held short-term public benefits outweighed by "the long-term detriments society may suffer." *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 196 (2d Cir. 2024).

II. Courts and regulators abroad show where the line falls. In *GEMA v. OpenAI*, a Munich court held that OpenAI's models reproduce protected works; in *GEMA v. Suno*, it held that training copies made in the United States were not fair use where the model reproduced the works. Korea, whose fair use statute weighs four factors that track Section 107, has advised that training on news articles to offer summaries without permission is unlikely to be fair use. The European Union ("EU"), Japan, and China condition unlicensed training. A fair use ruling here would not give U.S. developers the advantage the DOJ predicts.

III. The training-data market is real, and the Copyright Office has found it. More than 120 AI licensing agreements have been publicly reported, both Defendants have signed them, and the Copyright Clearance Center, whose licensing system decided factor four in *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913 (2d Cir. 1994), now licenses AI uses. Defendants even traded training data with each other, a market of their own that bypasses the rightsholders whose works make up those data sets. Their answer that rightsholders can simply opt out reaches only grounding, and it fails even there: an opt-out binds only crawlers that identify themselves, and a work already used in training cannot readily be removed.

IV. Each use (acquisition, training, grounding, and output) must be analyzed separately, on a record of what was used. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 533 (2023). The manner of acquisition bears on the character of the use, *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562-63 (1985); *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 477-79 (2d Cir. 2004), and a developer that cannot document what it copied cannot carry its burden on an affirmative defense, *see Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994).

On this record, the fair use factors favor Plaintiffs. The Court should grant Plaintiffs' motions for summary judgment and deny Defendants' motions.

ARGUMENT

I. NATIONAL SECURITY, PROPERLY WEIGHED UNDER FACTORS ONE AND FOUR, CUTS AGAINST UNPAID AND UNDOCUMENTED TRAINING DATA.

The fourth fair use factor asks not only whether a use harms the market for the original, but also what the public stands to gain. *See Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 35 (2021) (considering the public benefits the copying will likely produce). The Court asked whether those benefits are "related to copyright's concern for the creative production of new expression." *Id.* at 35. The United States has invited the Court to place national security on that scale, on the side of fair use. *See* Statement of Interest of the United States [ECF No. 1682] ("DOJ SOI") at 2, 18-19. To the extent national security belongs in the analysis at all, the DOJ's submission places it on the wrong side of the balance. Properly considered, the security interests of the United States favor a training-data ecosystem that is paid for and documented, not one built on uncompensated and unrecorded copying.

The Third Circuit recently rejected a similar argument. Affirming that an AI developer's unlicensed use of Westlaw headnotes to train its system was not fair use, the court held that "[w]hile some AI technology may implicate national security concerns, this does not give ROSS carte blanche to violate copyright law merely because it incorporates AI." *Thomson Reuters Enter. Ctr. GmbH v. ROSS Intel. Inc.*, No. 25-2153, slip op. at 28 (3d Cir. Sept. 29, 2026).

A. Professionally Produced Information Is Economic Infrastructure, and a Supply Chain That No One Pays For Will Not Survive.

Generative AI depends on a continuing supply of accurate, edited, professionally produced, originally created text and images: reporting, books, reference works, scholarship, and

illustration. That content is produced by people and institutions whose work is financed by the market for it. A fair use rule that allows the largest developers to take that human-made content without payment does not merely redistribute value; it degrades the input on which the models themselves depend. That result disserves both the purpose of copyright and the quality of generative AI itself.

Text is abundant: users generate it constantly, and models now generate more. What neither users nor models supply is the professional work that makes information reliable: editorial selection, original reporting, fact-checking, peer review, and correction of the published record. Publishers finance that work, along with the royalties that pay authors. Unlicensed training takes the value of that investment along with the expression it produces, a harm most acute in scholarly and educational publishing, where verification and peer review are the product.

The record shows that Defendants understand this dynamic. The News Plaintiffs cite a Microsoft document stating: "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: It is highly unusual that an end-product threatens the economic foundations of its essential suppliers, but that is the situation we have created for our LLM [large language model] business with respect to its 'content supply chain.'" News Plaintiffs' Combined Summary Judgment Brief [ECF No. 1977-1] ("News Pls.' Br.") at 2; *see also id.* at 73. Microsoft's Dr. Glen Weyl likewise described compensating creators as "in the best interests of my employer, of my country, and of many other groups I belong to." *Id.* at 74.

If the sources of high-quality text and images cannot sustain themselves, future models will be trained on less reliable material, and the public will receive less reliable outputs. Models

trained indiscriminately on model-generated content can suffer "irreversible defects." Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 Nature 755, 755 (2024). Sustaining that supply is a matter of economic resilience, which bears directly on the security interest the DOJ invokes. *See Oracle*, 593 U.S. at 35-36 (weighing the public benefits of the copying against the losses it causes under factor four). The Second Circuit has rejected a public-benefit defense of the kind the Defendants assert, holding that "short-term public benefits" were "outweighed not only by harm to Publishers and authors but also by the long-term detriments society may suffer." *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 196 (2d Cir. 2024).

The DOJ argues that licensing fees "would disproportionately benefit legacy media outlets," and that LLMs can help "level the playing field between mainstream and independent publishers." DOJ SOI at 4. Yet the DOJ acknowledges that publishers "could enter (and have entered) into licensing agreements" with developers for access to their content. *Id.* at 4 n.13. A supply chain that compensates its suppliers, large and small, is one that lasts. License payments are an exchange for value. Developers pay for every other input to their products, and nothing exempts copyrighted content from that rule. If anything, the DOJ's position would subsidize AI developers, which are among the most highly valued companies in the world. The factor-four inquiry asks whether unrestricted and widespread conduct of the sort engaged in by the defendant would substantially harm the market for the original. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994). The DOJ concedes that this inquiry "can be relevant," though it would confine it to the specific use and to substitutive similarity. DOJ SOI at 16 n.17. Applied to the information supply chain on which generative AI depends, the answer is yes, and the public shares in that harm.

B. The Government's Own Security Agencies Call for Documented Training Data.

A model is only as trustworthy as the data used to build it. A developer that cannot say what went into its training corpus cannot audit that corpus for data poisoning, for content placed by hostile foreign states, or for material engineered to manipulate model behavior. Provenance, meaning a documented record of what was collected and from where, is the precondition to any meaningful security review.

The government's own security agencies have applied this principle directly to AI training data. In joint guidance issued in May 2025, the National Security Agency ("NSA"), the Cybersecurity and Infrastructure Security Agency, and the Federal Bureau of Investigation identified the "data supply chain" as a significant security risk to AI systems, warned that training sets can be poisoned by malicious actors, and recommended that organizations track data provenance. *See* Nat'l Sec. Agency et al., *AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems*, at 1, 6, 8-10 (May 22, 2025) ("In short, the security of AI systems depends on thorough verification of training data and proactive measures to detect and prevent the introduction of inaccurate material"). Allied guidance issued in March 2026, led by Australia's cyber agency and co-sealed by the NSA and the agencies of six other nations, likewise tells organizations to "[c]hoose models with AIBOMs [AI bills of materials] or equivalent documentation that list their training data sources." Austl. Signals Directorate's Austl. Cyber Sec. Ctr. et al., *Artificial Intelligence and Machine Learning: Supply Chain Risks and Mitigations* 11 (Mar. 2026). Documentation and payment travel together. Every license identifies a work, a source, and a permitted use, so a developer that licenses its data gets a provenance record as a byproduct. A developer that copies without permission has every reason to keep none.

The record here illustrates the problem. Plaintiffs' evidence indicates that OpenAI acquired content, including through third-party sources such as Common Crawl, without checking whether it was paywalled or subject to terms of use, stripped the copyright notices that identified its owners, and abandoned the opt-out tool it had announced. News Pls.' Br. at 8-11, 10 n.2, 77. Microsoft, which shared portions of its Bing search index with OpenAI, argues that it would not matter "if the Bing Index contained works scraped from shadow libraries or other illicit sources." Def. Microsoft Corp.'s Mem. of Law in Supp. of Mot. for Summ. J., No. 23-cv-10211 [ECF No. 816] ("Microsoft Books Br.") at 23. A Microsoft scientist warned that an OpenAI output filter would amount to an "accidental cover up," leaving "people who have a right over the content having less visibility into what was used for training." News Pls.' Br. at 13. A corpus assembled this way is the opposite of the documented data supply chain the security guidance calls for, and the DOJ's theory would reward it with a fair use holding.

Amici do not ask the Court to decide questions of national security policy. They ask only that, if the Court accepts the DOJ's invitation, it weigh the whole of that interest. So weighed, it favors Plaintiffs.

C. National Security Also Weighs Against Fair Use Under the First Factor.

If the Court weighs national security, it should do so under the first factor as well, because the character of a use includes how the works were obtained. *See* Part IV, *infra*. Plaintiffs' evidence indicates that the works came from pirate sites, from a non-commercial license, and from search crawls that publishers permitted only for search. Class Pls.' Mem. of Law in Supp. of Mot. for Partial Summ. J. [ECF No. 1881] ("Class Pls.' Br.") at 5-6, 11; News Pls.' Br. at 8, 10-11, 57. Those paths left rightsholders unable to see, price, or refuse the uses made of their works. OpenAI's own expert states that "[n]o individual work or set of works is

required" for training. News Pls.' Br. at 58. "[U]nlike necessity, ease is not a justification for copying." *Thomson Reuters*, slip op. at 22.

II. COURTS AND REGULATORS ABROAD CONDITION AI TRAINING ON LAWFUL ACCESS AND RIGHTSHOLDER CHOICE, SO A FAIR USE RULING WOULD NOT GIVE U.S. DEVELOPERS THE ADVANTAGE THE DOJ PREDICTS.

The DOJ contends that a licensing outcome would disadvantage U.S. developers against foreign adversaries "not so encumbered." DOJ SOI at 2; *see id.* at 4 n.13. The decisions and legislation of other jurisdictions bear on that contention, and on whether a market for licensing training data is "traditional, reasonable, or likely to be developed." *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 930 (2d Cir. 1994). Amici offer them as context, not as authority binding on this Court.

On November 11, 2025, the Munich Regional Court I ruled against OpenAI on conduct closely related to that at issue here. *GEMA v. OpenAI*, Landgericht München I [Regional Court of Munich I], Az. 42 O 14139/24 (Nov. 11, 2025) (as summarized in Press Release No. 11, LG München I (Nov. 11, 2025) (unofficial English translation), https://ifrro.org/resources/documents/General/German_Court_OpenAI_Memory_Output_Infringe_Copyright_NOV25.pdf). The court held that the memorization of song lyrics within GPT-4 and GPT-4o is itself a reproduction of the protected works. *Id.* It further held that the German implementation of the EU text-and-data-mining exception, while it covers assembling a training set, does not extend to a model's retention of the works. *Id.* The court also held OpenAI, not its users, responsible for infringing outputs. *Id.*

That decision is on appeal, and amici do not suggest that German copyright law governs here. A second action, *GEMA v. Suno*, also on appeal, found copyright infringement for similar

uncompensated use. The Munich court held that, because the works reappeared in Suno's outputs, training copies made in the United States were "not covered by the fair use doctrine under 17 U.S.C. § 107," distinguishing *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025), and *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026 (N.D. Cal. 2025). *GEMA v. Suno*, LG München I, Az. 42 O 763/25 (July 31, 2026); *see* Press Release No. 16, LG München I (July 31, 2026) (amici's translation). Because the training copies were made in the United States, the court applied U.S. law to them.

A fair use jurisdiction has reached a similar view. Korea's Copyright Act directs courts to weigh four factors that track Section 107. Jeojakgwonbeop [Copyright Act] art. 35-5(2) (S. Kor.). Korea's culture ministry and copyright commission have advised that a developer that trains on full news articles without permission and sells a summary service is unlikely to make a fair use, because the use serves "the same or a very similar purpose as the publisher's original news articles." Ministry of Culture, Sports & Tourism & Korea Copyright Comm'n (S. Kor.), *Guide on Applicability of the Fair Use Doctrine to Training of Generative AI Models* 59 (Feb. 2026) (English ed. May 18, 2026). The guide describes its examples as illustrative, not authoritative. *Id.* at 3 (Preface).

Other jurisdictions also impose conditions. The EU's general text-and-data-mining exception reaches only "lawfully accessible works," and only where the use "has not been expressly reserved by their rightholders in an appropriate manner." Directive 2019/790, art. 4(1), (3), 2019 O.J. (L 130) 92, 113-14. Japan's broad exception does not apply where the use "would unreasonably prejudice the interests of the copyright owner," Chosakuken-hō [Copyright Act], Act No. 48 of 1970, art. 30-4 (Japan). And China requires generative AI providers to "[u]se data and foundational models that have lawful sources." Interim Measures for the Management of

Generative Artificial Intelligence Services art. 7(1) (promulgated by the Cyberspace Admin. of China et al., July 10, 2023, effective Aug. 15, 2023) (China Law Translate trans.).

The EU's AI legislation points in the same direction. Article 53 of the EU Artificial Intelligence Act requires providers of general-purpose AI models to adopt a policy to comply with EU copyright law, including rightsholders' reservations of rights, and to make publicly available a sufficiently detailed summary of the content used for training. Regulation (EU) 2024/1689, art. 53(1)(c)-(d), 2024 O.J. (L 2024/1689). The copyright-policy obligation applies "regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place." *Id.* recital 106. European law thus treats disclosure as a condition of market access, and disclosure is what makes licensing possible.

The licensing market abroad is concrete. GEMA, the German collecting society for composers, lyricists, and music publishers, has published a licensing model for generative AI uses of its repertoire. *See KI Hub*, GEMA, https://www.gema.de/en/about-gema/ki-hub (last visited Oct. 30, 2026) ("GEMA has developed a two-pronged model setting out a licensing framework for generative AI services"). That is the kind of evidence *Texaco* credited: an organized clearance mechanism, offered by an established collective, for the use at issue. *See* 60 F.3d at 930-31.

The DOJ's premise therefore does not hold. The nations with whom U.S. companies most directly compete and partner are converging on conditions of lawful access, rightsholder choice, or disclosure. A fair use ruling here would not free U.S. models from those conditions.

III. THE TRAINING-DATA MARKET IS REAL, AND THE OPT-OUT DEFENSE CANNOT SUBSTITUTE FOR IT.

A. The Market Exists, and the Defendants Participate in It.

The Class Plaintiffs have identified more than 120 publicly reported AI licensing agreements, and an industry-wide outlay, for licenses to use books in AI training and for compensation for prior use, "well-north of $1 billion." *See* Class Pls.' Br. at 37. Both Defendants have entered such agreements. *Id.*; News Pls.' Br. at 71-72. The Copyright Clearance Center, the same organization whose licensing system decided factor four in *Texaco*, 60 F.3d at 930-31, now offers aggregate licensing for AI uses. *See* Class Pls.' Br. at 37. Applying *Texaco*'s standard, the Copyright Office found that licensing markets "exist or are 'reasonable' or 'likely to be developed,' for certain copyright sectors, types of training or uses, and models," and that "[w]here licensing markets are available to meet AI training needs, unlicensed uses will be disfavored under the fourth factor." U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 70-71 (pre-publication version May 2025) (footnote omitted). Amici and their members participate in that market as licensors.

The Third Circuit has recognized the emerging training-data market in deciding whether AI training is fair use. It found that "the evidence shows that the market for licensing headnotes as text to train AI is rapidly developing." *Thomson Reuters*, slip op. at 27. There, the court held that ROSS's unlicensed use of Thomson Reuters' works usurped Thomson Reuters' opportunity to enter the derivative market of training-data licensing. *See id.* The court distinguished generative models in assessing transformation, *id.* at 18 n.7, but its market holding rested on the existence of a training-data market, which is equally present here.

Defendants know this market well and have avoided paying into it for the works at issue here. The public record shows Defendants trading training data with each other. Microsoft

provided OpenAI with access to the Bing search index ("Project Taxi"), operated a web crawler on OpenAI's behalf ("Project Mango"), and received payment and OpenAI training data in return. *See* News Pls.' Br. at 10, 31-33. OpenAI's Chief Research Officer characterized the Bing Index deal as "a trade." *Id.* at 32. Training data, in other words, has value that Defendants themselves recognize, price, and exchange with one another, while the creators whose works make up those data sets receive nothing.

B. The Circularity Objection Fails Where, as Here, There Are Real Buyers, Real Sellers, and a Clearance Mechanism.

Defendants respond that a copyright owner is not entitled to a license fee for a transformative use, and that treating the loss of such a fee as market harm would make the fourth factor circular. *See* Defs.' Mem. of Points & Authorities in Supp. of Mot. for Summ. J., No. 23-cv-11195 [ECF No. 1496] ("OpenAI News Br.") at 11 (citing *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614 (2d Cir. 2006); *Kadrey*, 788 F. Supp. 3d at 1052; *Bartz*, 787 F. Supp. 3d at 1032).

Texaco answers that concern. The Second Circuit acknowledged that a copyright holder can always point to an unpaid fee, so it limited cognizable harm to "traditional, reasonable, or likely to be developed markets." 60 F.3d at 929-30 & n.17. Publishers built a workable licensing market through the Copyright Clearance Center, and major corporations, including Texaco, were paying into it. The court therefore held that lost licensing revenue counted under the fourth factor. *Id.* at 930-31. As the court explained, "[t]he vice of circular reasoning arises only if the availability of payment is conclusive against fair use." *Id.* at 931.

The harm is cognizable because the market for training data already exists and continues to develop. It is a body of executed transactions. *See* Part III.A, *supra*. As noted, the Third Circuit treated that market as cognizable even though the rightsholder had not yet licensed into it.

Thomson Reuters, slip op. at 27. *Kadrey* and *Bartz* do not bind this Court, and neither had a record like this one. *Kadrey* itself observed that markets for news articles "might be even more vulnerable to indirect competition from AI outputs," 788 F. Supp. 3d at 1059, and *Bartz* noted that if the models had produced infringing outputs, "this would be a different case," 787 F. Supp. 3d at 1032.

Nor does *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), counsel otherwise. The court there held that snippet view did not supply a meaningful substitute for the books. *Id.* at 224. That conclusion turned on a display feature built to reveal only fragments. Market substitution also does not require verbatim copying. A response that summarizes or synthesizes a book, article, or subscription product can satisfy a user's need for the original without reproducing an identifiable passage. As the Copyright Office explained, "[a] user for whom the augmented response 'satisf[ies] the . . . need' for the original work will not pay to obtain it in the marketplace." *Part 3*, *supra*, at 63-64 (footnote omitted); *see id.* at 65 ("The speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data.").

Microsoft argues that licensing is impracticable because the works belong to millions of dispersed authors and cannot readily be priced. *See* Microsoft Books Br. at 28-29. But that is the problem collective licensing exists to solve, and it is the problem the Copyright Clearance Center solved in *Texaco*. *See* 60 F.3d at 930-31. Collective licensing also reaches the long tail of works that fall outside the direct licenses AI developers are already signing. Microsoft relies on *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1279 (11th Cir. 2014). Microsoft Books Br. at 28. But *Patton* itself approved weighing the availability of licenses under factor four. 769 F.3d at 1279. Defendants' transaction costs are also largely self-imposed. The chief obstacle to pricing

and clearing the works is that the training copies carried almost none of the information identifying their owners. *See* News Pls.' Br. at 77; Part I.B, *supra*; Part III.C, *infra*.

C. Opacity Suppresses the Market, and an Opt-Out Cannot Cure It.

A rightsholder cannot license what it cannot see. OpenAI relied on Common Crawl, a third party Plaintiffs never authorized, and, Plaintiffs contend, so sidestepped publishers' robots.txt instructions. News Pls.' Br. at 9-10. It obtained the New York Times Annotated Corpus under a license limited to non-commercial use, which Plaintiffs contend it secured by misrepresenting its commercial aims, and used it for commercial training. *Id.* at 10-11, 57. Microsoft took content that, Plaintiffs' evidence indicates, publishers had allowed it to crawl only for Bing search and passed it to OpenAI for model development and training. *Id.* at 8, 57. And OpenAI announced, then abandoned, a "Media Manager" tool that would have let creators opt out of training. *Id.* at 10 n.2. Each practice deprived rightsholders of the information needed to identify, price, and license the use of their works.

The same point answers Defendants' reliance on robots.txt files and meta tags, which they invoke as an implied license for Browse and Copilot grounding and as proof that publishers could have protected themselves. *See* OpenAI News Br. at 14-15, 20 & n.15; Def. Microsoft Corp.'s Mem. of Law in Supp. of Mot. for Summ. J., No. 24-cv-04872 [ECF No. 729] at 8-9, 31-34. As a threshold matter, copyright is a permission-based system: the Copyright Act gives owners the exclusive right to authorize reproduction, 17 U.S.C. § 106(1), and does not make that right depend on an owner's objection. An opt-out mechanism works only if the rightsholder knows who is crawling and for what purpose. When a search crawler's output is repurposed for training, or a crawler is operated by one company on behalf of another, as the Project Mango record shows, a robots.txt instruction directed to one purpose or one agent cannot be an informed choice about another. Publishers have opted out where they could: among the most critical web

domains, the share of news-site tokens fully blocked by robots.txt rose from 3% in 2023 to nearly 45% by April 2024. Shayne Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons*, *in* 37 Advances in Neural Information Processing Systems 108042, 108047 (2024). But those restrictions operate only going forward and only against crawlers that identify themselves. Nor can a later opt-out undo training that has already occurred. Rightsholders told the Copyright Office that, to the best of their knowledge, "technologies do not yet exist that can effectively remove entire works at scale from an AI model after it has been trained," though they "might be coming." U.S. Copyright Office, *Part 3*, *supra*, at 101 & n.568 (quoting Copyright Alliance Initial Comments at 70). And a court in this District has rejected the argument that a publisher's failure to deploy robots.txt impliedly licenses a use. *See Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 563 (S.D.N.Y. 2013).

D. Opacity Entrenches Incumbents; Disclosure and Licensing Promote Competition.

The DOJ contends that licensing would favor the largest technology companies and "disproportionately benefit legacy media outlets." DOJ SOI at 4. Amici's experience, as rightsholders, their representatives, and participants in this market, is the opposite. Few developers can absorb the legal risk of copying at scale or trade access to proprietary indices and crawlers among themselves, as the Project Taxi and Project Mango records show. Smaller developers can compete by acquiring documented, licensed data with known provenance and known terms, and independent rightsholders participate through direct and collective licensing. A fair use rule that rewards undocumented copying would lock in the advantage of the few firms able to engage in it.

IV. THE COURT SHOULD ANALYZE EACH USE SEPARATELY, ON A RECORD OF WHAT WAS USED, AND SHOULD NOT REWARD DEVELOPERS FOR KEEPING NO RECORDS.

First, fair use must be assessed use by use. *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 533 (2023) ("The same copying may be fair when used for one purpose but not another."); *Authors Guild*, 804 F.3d at 216-24 (analyzing search and snippet view as distinct uses). Acquisition, training, grounding, and output each involve the reproduction of Plaintiffs' works, and each must be justified on its own; liability does not depend on whether the outputs infringe. That analysis requires the Court to consider a record of what was used for each. The DOJ concedes that certain output uses "may not be transformative if the LLM reconstructs and disseminates an original copyrighted work," DOJ SOI at 11, and Plaintiffs' evidence shows Defendants' grounding features delivering the text of current news articles to users, News Pls.' Br. at 13-20. A commercial service that retrieves and delivers a rightsholder's content on demand is not fair use, *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 176-81 (2d Cir. 2018), and courts have reached the same conclusion for news excerpting and clipping services, *Meltwater*, 931 F. Supp. 2d at 552-54; *Pac. & S. Co. v. Duncan*, 744 F.2d 1490, 1496-97 (11th Cir. 1984). Training cannot be walled off from those outputs here: OpenAI's Vice President of Research explained that "[w]e train our networks to memorize the training data," and OpenAI's own expert found 409,219 verbatim 25-word sequences from 3,191 Times articles in 19,330 ChatGPT conversations drawn from a 20-million-conversation sample, outputs Plaintiffs attribute to training and grounding. News Pls.' Br. at 12, 15 & n.4.

Second, the manner in which works were obtained bears on the character of the use, the first factor in a fair use analysis. *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562-63 (1985) (stating that "the propriety of the defendant's conduct" is "relevant to the 'character' of the use") (quoting 3 Nimmer on Copyright § 13.05[A]); *NXIVM Corp. v. Ross*

Inst., 364 F.3d 471, 477-79 (2d Cir. 2004) (bad faith, including whether the defendant knew its access "was unauthorized or was derived from a violation of law or breach of duty," bears on the first factor). Courts have questioned how much weight bad faith deserves, *see NXIVM*, 364 F.3d at 478-79; *Oracle*, 593 U.S. at 32-33, but the Third Circuit recently held that, "to the extent good faith remains relevant," an AI developer's bad faith weighs against fair use. *Thomson Reuters*, slip op. at 22 n.9. Acquisition is also a use in its own right: pirating copies "to build a research library without paying for it, and to retain copies should they prove useful for one thing or another, was its own use—and not a transformative one." *Bartz*, 787 F. Supp. 3d at 1026. Nor is public availability permission. As the Copyright Office has observed, "'[p]ublicly available' is not synonymous with 'authorized.'" U.S. Copyright Office, *Part 3*, *supra*, at 13 n.70. The search crawls described above illustrate the point. *See* News Pls.' Br. at 8, 57. The Court should consider how Defendants acquired their training data. And because fair use is an affirmative defense, a developer that cannot document what it copied, from where, and on what terms cannot carry its burden of proving that each use was fair. *See Campbell*, 510 U.S. at 590.

CONCLUSION

Amici respectfully ask the Court to hold that (1) each use of Plaintiffs' works (acquisition, training, grounding, and output) must be analyzed separately, on a record that identifies what was used; (2) the manner of acquisition, including a developer's failure to keep or disclose a record of provenance, bears on the character of the use; (3) as Plaintiffs show, the second and third factors point the same way; (4) the market for licensed training data is a cognizable market under *Texaco*, and harm to it weighs against fair use; and (5) to the extent the Court considers it, the national security interest the DOJ invokes weighs against fair use as well.

The Court should grant Plaintiffs' motions for summary judgment and deny Defendants' motions.

Dated: October 30, 2026
Respectfully submitted,

/s/ Tod Cohen
Tod Cohen (*pro hac vice* motion pending, ECF No. 2115)
LAW OFFICE OF TOD COHEN
1159 Church Street
San Francisco, CA 94114
(408) 410-8329
tcohen@gmail.com
Counsel for Amici Curiae

CERTIFICATE OF COMPLIANCE

I certify that this brief contains 5,670 words, excluding the caption, table of contents, table of authorities, signature block, and this certificate, as counted by the word-processing program used to prepare it, and complies with the 6,000-word limit set by the Court's Order of August 7, 2026 (ECF No. 1663), and with Local Civil Rule 7.1(c).

/s/ Tod Cohen
Tod Cohen