UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION	No. 25-md-3143 (SHS) (OTW) Hon. Sidney H. Stein Hon. Ona T. Wang

BRIEF OF AMICI CURIAE CENTER FOR AI AND DIGITAL POLICY, AI EXPERTS, AND LEGAL SCHOLARS IN CONNECTION WITH THE PARTIES' MOTIONS FOR SUMMARY JUDGMENT

TABLE OF CONTENTS

TABLE OF CONTENTS 2

TABLE OF AUTHORITIES 4

INTEREST OF AMICI CURIAE 8

SUMMARY OF ARGUMENT 13

ARGUMENT 16

I. EFFECTIVE AI GOVERNANCE REQUIRES TRANSPARENCY CONCERNING TRAINING DATA 16

A. Training-data provenance is a foundational accountability mechanism. 16

B. Documentation enables meaningful legal and regulatory review. 18

C. Established AI governance frameworks recognize transparency and traceability as core requirements. 19

II. COPYRIGHT LAW SHOULD NOT REWARD OPACITY IN AI DEVELOPMENT 22

A. Fair use requires identification of the particular use. 22

B. Source and lawful access can bear on the character of the use. 24

C. Poor recordkeeping should not create an evidentiary advantage. 25

III. TRAINING-DATA DOCUMENTATION PROTECTS RIGHTS AND SUPPORTS RESPONSIBLE INNOVATION 27

A. Documentation can coexist with protection for trade secrets and proprietary information. 27

B. Documentation supports licensing, auditing, compliance, and dispute resolution. 28

C. Clear accountability rules provide certainty for responsible developers. 30

IV. THE COURT SHOULD APPLY EXISTING COPYRIGHT LAW CONSISTENTLY WITH AI ACCOUNTABILITY 31

A. Human responsibility remains essential when automated systems operate at scale 31

B. General-purpose AI developers should preserve sufficient evidence of training-data provenance and use 33

C. The Court should avoid an incentive structure that rewards opacity. 34

CONCLUSION 35

CERTIFICATE OF COMPLIANCE 37

TABLE OF AUTHORITIES

CASES

Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023) 13, 22

Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) ("Google Books") ...15, 23

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) 22

Google LLC v. Oracle America, Inc., 593 U.S. 1 (2021).....22

Harper & Row, Publishers, Inc. v. Nation Enterprises, 471 U.S. 539 (1985)..15, 24

NXIVM Corp. v. Ross Institute, 364 F.3d 471 (2d Cir. 2004) 15, 24

STATUTES

17 U.S.C. § 107 .....passim

Cal. Civ. Code § 3111. 26

OTHER AUTHORITIES

Abeba Birhane et al., *SoK: AI Auditing: The Broken Bus on the Road to AI Accountability*, 2d IEEE Conf. on Secure & Trustworthy Machine Learning (2024) 10, 24

Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (2022) 10, 26

Anu Bradford, *The False Choice Between Digital Regulation and Innovation,* 119 Nw. U. L. Rev. 377 (2024).....20, 28

Center for AI and Digital Policy, *Complaint and Request for Investigation, Injunction, and Other Relief*, In re OpenAI, Inc., Federal Trade Commission (Mar. 30, 2023).....8

Hachette Book Group, Inc. v. Internet Archive, 115 F.4th 163 (2d Cir. 2024).....21

Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* 140–47 (Stanford Univ. Press 2010).....24

Lorraine Kisselburgh & Jonathan Beever, *The Ethics of Privacy in Research and Design: Principles, Practices, and Potential*, in *Modern Socio-Technical Perspectives on Privacy* 395-426 (Springer 2022).....19

Marc Rotenberg & Eleni Kyriakides, *The AI Policy Sourcebook* (CAIDP 2025) ..18

Margaret Mitchell et al., *Model Cards for Model Reporting*, Proc. Conf. Fairness, Accountability & Transparency 220 (2019).....12, 15

Meredith Broussard, *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech* (MIT Press 2023).....17

Nathalie A. Smuha, *Algorithmic Rule By Law: How Algorithmic Regulation in the Public Sector Erodes the Rule of Law* (2025) 10, 19

National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (2023) 18

Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).....9

The Public Voice, *Universal Guidelines for Artificial Intelligence*, Guidelines (Oct. 23, 2018).....18

Timnit Gebru et al., *Datasheets for Datasets*, 64 Commc'ns ACM 86 (2021).12, 15

U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 36 (2025) ("Copyright Office Report").....passim

Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* 47–106 (2019).....9, 15, 29

Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 Science 842, 842–45 (2024).....19

RULES

Fed. R. Civ. P. 56(c).....17, 23, 32

FOREIGN AND INTERNATIONAL AUTHORITIES

Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225 (Sept. 5, 2024).....18

Council of Europe, *North Macedonia Joins the Council of Europe AI Convention* (May 2026).....18

OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (2019, amended 2024).....18

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1 18

UNESCO, Recommendation on the Ethics of Artificial Intelligence (Nov. 23, 2021).....18

INTEREST OF AMICI CURIAE

The Center for AI and Digital Policy ("CAIDP") is an independent nonprofit research and educational organization that promotes artificial intelligence policies grounded in fundamental rights, democratic institutions, and the rule of law. CAIDP conducts comparative research on national AI policies, advises governments and international organizations, trains AI policy professionals, and participates in legal and policy proceedings concerning artificial intelligence. CAIDP has no financial interest in the outcome of this litigation.

This brief is submitted on behalf of CAIDP and individual AI experts and legal scholars (affiliations are for identification only).

INDIVIDUAL AMICI

Ricardo Baeza-Yates, Professor, Universitat Pompeu Fabra, Barcelona, Spain

Meredith Broussard, Ph.D., Professor, New York University; Research Director, Arthur L. Carter Journalism Institute

Nita Farahany, Robinson O. Everett Professor of Law and Philosophy, Duke Law School

Gry Hasselbalch, Ph.D., Academic Director and Co-Founder, DataEthics.eu

Merve Hickok, Lecturer, University of Michigan, School of Information; Author, *"From Trustworthy AI Principles to Public Procurement Practices"*

Syed Ishtiaque Ahmed, Associate Professor of Computer Science, University of Toronto

Lorraine Kisselburgh, Ph.D., lecturer and retired professor, School of Interdisciplinary Studies, Purdue University

Marcos López Oneto, Ph.D., Lecturer at the Innovation and Artificial Intelligence Laboratory (IALAB), University of Buenos Aires, Argentina.

Valsamis Mitsilegas, Professor of European and Global Law, University of Liverpool

Pablo G. Molina, Ph.D., Faculty Lecturer, Georgetown University; Executive Director, International Applied Ethics and Technology Association

Helen Nissenbaum, Andrew H. & Ann R. Tisch Professor of Information Science, Cornell Tech; Director of the Digital Life Initiative

Gabriela Ramos, former Assistant Director-General for Social and Human Sciences, UNESCO; former G20 and G7 Sherpa, OECD

Carolina A. Rossini, Professor of Practice, UMass School of Public Policy; Director for Programs, Public Interest Technology Initiative @UMass

Stuart Russell, Distinguished Professor Emeritus, Computer Science Division, University of California, Berkeley

Edward Santow, Industry Professor, Responsible Technology, University of Technology Sydney

Ingrid Schneider, Professor of Political Science, Department of Informatics, University of Hamburg, Germany

John Shattuck, President Emeritus, Central European University

Nathalie A. Smuha, Assistant Professor of Law, University of Toronto Faculty of Law; former Coordinator, European Commission High-Level Expert Group on Artificial Intelligence

Hannes Werthner, University Professor (ret.), Technical University Vienna

CAIDP has addressed the transparency and accountability obligations of developers of general-purpose artificial intelligence systems before federal agencies and courts. In March 2023, CAIDP filed a complaint with the Federal Trade Commission concerning OpenAI's deployment of GPT-4. The complaint emphasized independent assessment, transparency, accountability, and effective safeguards for commercial AI systems. *See* Center for AI and Digital Policy, *Complaint and Request for Investigation, Injunction, and Other Relief*, In re OpenAI, Inc., Federal Trade Commission (Mar. 30, 2023). CAIDP later submitted additional materials concerning OpenAI and participated in proceedings before the U.S. Copyright Office and the National Telecommunications and Information Administration concerning generative AI and AI accountability.

CAIDP has also addressed training-data transparency directly. In *X.AI LLC v. Bonta*, No. 26-1591 (9th Cir.), CAIDP filed an amicus brief supporting California's Generative Artificial Intelligence Training Data Transparency Act,

Cal. Civ. Code § 3111. The statute requires developers of public-facing generative AI systems to publish specified information about training datasets, including sources, types of data, intellectual-property status, processing, and the use of synthetic data. CAIDP explained that these disclosures permit regulators, researchers, rights-holders, and the public to evaluate how generative AI systems were developed and whether their data practices are lawful and accountable.

CAIDP's Global Academic Network includes scholars in computer science, law, data governance, human rights, public policy, and technology accountability. Their work informs CAIDP's understanding of the governance questions presented here. Professor Stuart Russell, a CAIDP board member and a 2023 CAIDP AI Policy Leader, has emphasized that the design and deployment of advanced AI systems must remain subject to human control and institutional responsibility. Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019). Professor Virginia Dignum has developed a lifecycle approach to responsible AI that places responsibility on the people and organizations that design, develop, deploy, and use AI systems. Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* 47–106 (2019).

Other members of the Global Academic Network have examined related aspects of accountability. Professor Abeba Birhane and her coauthors have shown that AI auditing depends on access to information, meaningful institutional

independence, and mechanisms that connect evaluation to accountability. Abeba Birhane et al., *SoK: AI Auditing: The Broken Bus on the Road to AI Accountability*, 2d IEEE Conf. on Secure & Trustworthy Machine Learning (2024). Professor Alessandro Mantelero has developed a human-rights impact-assessment framework for AI that depends on structured evaluation of data, context, affected interests, and risks. Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (2022). Professor Nathalie Smuha has examined the relationship between algorithmic governance and the rule of law, emphasizing that technological systems must remain subject to legal standards and institutional review. Nathalie A. Smuha, *Algorithmic Rule By Law: How Algorithmic Regulation in the Public Sector Erodes the Rule of Law* (2025).

These authorities do not resolve the copyright claims. The Copyright Act supplies the governing rule. *They do explain why data provenance and documentation are central to responsible AI development and why legal rules should preserve, rather than weaken, incentives to maintain reliable records about training data.*

No party or counsel for a party authored this brief in whole or in part. No party or counsel for a party contributed money intended to fund the preparation or submission of this brief. No person other than the amici curiae or their members contributed money intended to fund the preparation or submission of this brief.

SUMMARY OF ARGUMENT

Generative artificial intelligence requires data. General-purpose language models often obtain very large quantities of data. The scale of those systems changes engineering practice. It does not eliminate legal responsibility for how data are acquired, copied, processed, retained, and used.

This litigation concerns several acts involving copyrighted works. Those acts include obtaining works, assembling and processing datasets, copying works during model development, retaining or transferring data, training particular model versions, and generating outputs. The Copyright Office has concluded that "different uses during AI development and deployment require separate consideration." U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 36 (2025) ("Copyright Office Report"). That conclusion follows the Supreme Court's instruction in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith,* 598 U.S. 508, 533 (2023), to identify the particular use alleged to be infringing before applying the fair-use factors.

OpenAI's own summary-judgment submission illustrates the point: it describes dataset construction, pretraining, evaluation, post-training, and deployment, and states that Library Genesis compilations were used to train GPT-3 and GPT-3.5 but not other models at issue. OpenAI Br. at 3, 5–9. Those model-specific facts underscore the value of provenance records.

CAIDP addresses a narrower question than the ultimate fair-use determination. The Court should not interpret copyright law in a way that gives developers an advantage from failing to know, preserve, or disclose facts about training-data provenance. Section 107 of the Copyright Act, 17 U.S.C. § 107, requires a fact-specific inquiry. That inquiry cannot operate reliably if legally significant facts disappear into the scale or complexity of an AI development process.

First, training-data provenance is a basic accountability mechanism. Technical researchers developed datasheets, model cards, audit methods, and related documentation practices because complex machine-learning systems cannot be evaluated from outputs alone. See Timnit Gebru et al., *Datasheets for Datasets*, 64 Commc'ns ACM 86 (2021); Margaret Mitchell et al., *Model Cards for Model Reporting*, Proc. Conf. Fairness, Accountability & Transparency 220 (2019). CAIDP's Global Academic Network has developed related work on responsible AI, impact assessment, auditing, contextual integrity, and the rule of law. The common premise is practical: meaningful review requires records that permit reviewers to reconstruct relevant decisions and data flows.

Second, copyright law does not treat "AI training" as a single legal event. *Warhol* requires identification of the use. The Copyright Office has applied that principle to generative AI and concluded that initial pretraining, subsequent training, retrieval-augmented generation, and other acts can require separate

analysis. Copyright Office Report at 36–37. *Authors Guild v. Google, Inc.,* 804 F.3d 202 (2d Cir. 2015) ("Google Books"), likewise distinguished among digitization, search functionality, snippet display, and distribution. The Court should assess the acts presented by the record rather than allowing the label "training" to resolve the analysis.

Third, source and lawful access can bear on the character of a use. The Copyright Office concluded that knowing use of a dataset consisting of pirated or illegally accessed works should weigh against fair use, although it is not determinative. Copyright Office Report at 52. That conclusion is consistent with *Harper & Row, Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 562–63 (1985), and with the Second Circuit's recognition in *NXIVM Corp. v. Ross Institute*, 364 F.3d 471, 482 (2d Cir. 2004), that bad faith does not automatically preclude fair use. Provenance therefore does not establish liability by itself. It supplies facts relevant to the statutory inquiry.

Fourth, inadequate documentation should not become an evidentiary benefit. A developer that maintains records may be able to identify whether a work entered a particular dataset, where it came from, how it was processed, and which models used it. A developer that does not maintain comparable records may be unable to answer those questions. Copyright doctrine should not place the first developer in a worse legal position simply because it preserved better evidence.

Fifth, documentation can coexist with legitimate trade-secret and security interests. Accountability does not require publication of model weights, source code, proprietary filtering methods, or every data record. Courts can use protective orders. Developers can retain detailed internal records while making more limited public disclosures. California's training-data transparency law follows that general model. The distinction between preserving information and publishing all information is essential.

CAIDP therefore asks the Court to apply existing copyright law in a manner that preserves incentives for reliable training-data records. The Court need not create a new cause of action or impose a technical documentation code. It can hold that the fair-use inquiry applies to identified acts; that provenance and access may be relevant where the governing factor makes them relevant; and that uncertainty created by a party's own inadequate records does not become affirmative evidence in favor of fair use.

ARGUMENT

I. EFFECTIVE AI GOVERNANCE REQUIRES TRANSPARENCY CONCERNING TRAINING DATA

A. Training-data provenance is a foundational accountability mechanism.

A general-purpose AI system is the product of organizational choices. Developers select data sources. They determine acquisition methods. They establish filters, deduplication procedures, quality criteria, and exclusion rules. They combine datasets. They decide which model versions use which datasets.

They retain or delete intermediate copies. They conduct post-training and evaluation. These decisions remain human and institutional decisions even when software automates portions of the process.

Training-data provenance records those decisions in a form that permits later review. Depending on the system, relevant records can identify the source of material, the date or period of collection, the method of acquisition, applicable licenses or restrictions, dataset versions, filtering and processing steps, and the models for which the data were used. The required level of detail can vary with context. The core function does not. Provenance connects an AI system to the data and decisions that produced it.

The machine-learning community developed documentation practices because undocumented datasets create predictable technical and governance failures. Gebru and her coauthors proposed "datasheets for datasets" to record motivation, composition, collection process, preprocessing, uses, distribution, and maintenance. Gebru et al., supra, at 86–92. Mitchell and her coauthors proposed model cards to provide standardized information about intended uses, evaluation, performance, and limitations. Mitchell et al., supra, at 220–29. These approaches do not assume that a final model can explain its own development history. They require organizations to preserve information during development.

The work of CAIDP Global Academic Network members reinforces this point. Dignum treats responsibility as a lifecycle obligation that attaches to the

people and organizations surrounding an AI system. Dignum, supra, at 47–106. Birhane and her coauthors explain that auditing cannot function as accountability when auditors lack access, independence, or institutional mechanisms for response. Birhane et al., supra. Mantelero's impact-assessment framework similarly begins with structured information about a system, its context, affected interests, and foreseeable effects. Mantelero, supra. These approaches differ in method and subject. Each depends on records that make evaluation possible.

B. Documentation enables meaningful legal and regulatory review.

Courts resolve copyright disputes through evidence. Section 107 directs courts to examine the purpose and character of the use, the nature of the copyrighted work, the amount used, and the effect on the potential market. 17 U.S.C. § 107. Each factor can depend on facts about acquisition, copying, processing, retention, and downstream use.

The current record shows why documentation is necessary. OpenAI states that GPT-3 and GPT-3.5 used Library Genesis compilations and that the other models at issue did not. OpenAI Br. at 3. The distinction is model specific. It connects a source collection to particular model versions. A court can evaluate the legal significance of that fact only because the development history can be reconstructed with some specificity.

Birhane's work on AI auditing identifies a similar institutional problem: review cannot produce accountability when evaluators lack the information needed

to test claims. Broussard likewise shows that automated systems reflect human and institutional choices and therefore require examination of the practices surrounding their design and deployment. Meredith Broussard, *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech* 8–23 (MIT Press 2023). These observations do not alter Rule 56 or copyright doctrine. They explain why a legal rule that rewards informational gaps would undermine reliable adjudication.

CAIDP has consistently taken that position in its own work. Its 2023 FTC complaint concerning GPT-4 emphasized that independent assessment of a powerful commercial AI system requires meaningful information about system design and operation. CAIDP's amicus filing *in X.AI v. Bonta* defended a statute requiring basic information about training sources and processing. The Universal Guidelines for AI likewise connect transparency to accountability and human responsibility. These instruments do not supply the governing copyright rule. They demonstrate the practical importance of preserving facts that legal institutions need to apply existing rules.

C. Established AI governance frameworks recognize transparency and traceability as core requirements.

Major AI governance frameworks use different terminology, but they converge on documentation, transparency, traceability, risk management, and accountability. The Universal Guidelines for Artificial Intelligence call for transparency, human determination, assessment, and accountability. The Public Voice, *Universal Guidelines for Artificial Intelligence*, Guidelines 1, 2, 5, 7 (Oct.

23, 2018). The OECD AI Principles §§ 1.3, 1.5; UNESCO Recommendation on AI Ethics ¶¶ 37-46; NIST AI Risk Management Framework (AI RMF 1.0), at 15, 22-23, European Union AI Act arts. 53(1)(a)–(b), Annex XII § 2(c); and the Council of Europe Framework Convention arts. 8-9 likewise require or promote forms of documentation and oversight. *See generally* Marc Rotenberg & Eleni Kyriakides, *The AI Policy Sourcebook* (CAIDP 2025).

The Council of Europe Framework Convention is especially relevant to CAIDP's rule-of-law concern. The Convention establishes principles concerning transparency and oversight, accountability and responsibility, reliability, and procedural safeguards. Secretary General Alain Berset, a 2026 CAIDP AI Policy Leader, has described the Convention as a legal framework for ensuring that AI development remains consistent with human rights, democracy, and the rule of law. See Council of Europe, *North Macedonia Joins the Council of Europe AI Convention* (May 2026). The Convention does not determine copyright liability. It reflects the broader principle that advanced AI systems should remain subject to institutional oversight grounded in law. Complexity strengthens rather than diminishes the need for documentation. AI governance scholarship has accordingly emphasized not only transparency and accountability but also intelligibility, traceability, and clear provenance of the data and actors involved. See Lorraine Kisselburgh & Jonathan Beever, *The Ethics of Privacy in Research and Design:*

Principles, Practices, and Potential, in *Modern Socio-Technical Perspectives on Privacy* 395-426 (Springer 2022).

Professor Smuha's scholarship develops the same rule-of-law point in a different setting. Algorithmic systems can weaken legal protections when institutions substitute technical process for legal accountability. Smuha, supra. The appropriate response is not to treat technical complexity as an exception from legal standards. It is to preserve the information and review mechanisms necessary to apply those standards.

Professor Bengio, a 2026 CAIDP AI Policy Leader, and Professor Russell have likewise emphasized that advanced AI requires both technical safeguards and governance. Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 Science 842, 842–45 (2024). Their concern focuses substantially on severe AI risks, not copyright. But the institutional lesson is general: technical development and governance operate together. Reliable governance depends on information about how systems are built and used.

California's AB 2013 translates this concept into a concrete training-data rule. The statute requires developers of public-facing generative AI systems to disclose categories of information concerning sources, data types, intellectual-property status, collection, processing, and synthetic data. It does not require publication of source code or model weights. The statute reflects a policy judgment

that basic information about training data is necessary for accountability while legitimate confidential interests can remain protected.

Professor Anu Bradford, a CAIDP board member and a 2024 AI Policy Leader, has explained why regulation and innovation should not be treated as inherently opposed. Anu Bradford, *The False Choice Between Digital Regulation and Innovation,* 119 Nw. U. L. Rev. 377 (2024). Clear legal obligations can create certainty and reward firms that invest in compliance. Documentation rules illustrate that point. They impose costs, but they also permit developers to prove lawful practices, respond to claims, negotiate licenses, and operate across jurisdictions.

II. COPYRIGHT LAW SHOULD NOT REWARD OPACITY IN AI DEVELOPMENT

A. Fair use requires identification of the particular use.

Section 107 requires a context-specific inquiry. The Supreme Court has repeatedly rejected categorical substitutes for the statutory factors. *Campbell v. Acuff-Rose Music, Inc.,* 510 U.S. 569, 577–85 (1994); *Google LLC v. Oracle America, Inc.*, 593 U.S. 1, 18–21 (2021); *Warhol*, 598 U.S. at 525–33. The first step is identifying the use that is alleged to infringe.

Warhol is particularly instructive. The Supreme Court explained that the same copyrighted work can be used for different purposes and that the analysis turns on the challenged use, not an abstract description of the secondary work. 598 U.S. at 533. The Copyright Office has applied that instruction to generative AI. It

concluded that copyrighted works are used in different ways during development and deployment and that those uses require separate consideration. *Copyright Office Report* at 36.

OpenAI's own technical description confirms the existence of distinct acts. It describes collecting and processing data, constructing training datasets, pretraining models, evaluating them, conducting post-training, and deploying systems. OpenAI Br. at 5–9. Plaintiffs challenge conduct at more than one point in that sequence. The Court should therefore identify the act under review before applying the statutory factors.

Second Circuit precedent supports that approach. In *Google Books*, the court distinguished the creation of digital copies, the development of search functionality, the display of snippets, and the distribution of copies to libraries. 804 F.3d at 216–29. The decision did not establish that any computational use of books is fair. It evaluated the particular functions and safeguards before the court.

Hachette Book Group, Inc. v. Internet Archive, 115 F.4th 163 (2d Cir. 2024), confirms that superficially similar technologies can receive different fair-use treatment when their purposes and market effects differ. The Court should likewise distinguish among acquisition, dataset construction, model training, retention, and outputs rather than treating "AI training" as a single use. The legal significance of each act depends on the claim, the record, and the applicable fair-use factor.

B. Source and lawful access can bear on the character of the use.

The Copyright Office has directly addressed lawful access. It concluded that knowing use of a dataset consisting of pirated or illegally accessed works should weigh against fair use without being determinative. Copyright Office Report at 52. It also cautioned against making lawful access a per se rule. Id. That is a balanced approach and consistent with existing authority.

Harper & Row treated the knowing exploitation of a purloined manuscript as relevant to the character of the use. 471 U.S. at 562–63. NXIVM held that bad faith does not automatically preclude fair use. 364 F.3d at 482.

This distinction is important here because OpenAI identifies Library Genesis as a source used for GPT-3 and GPT-3.5. OpenAI Br. at 3. CAIDP does not ask the Court to determine liability from the name of a repository. Nor does CAIDP ask the Court to infer facts not established in the summary-judgment record. The point is narrower: when a party's own evidence identifies the source of training material, the Court may consider any legally relevant characteristics of that source under the proper statutory factor.

A rule that source is categorically irrelevant would create an unnecessary conflict with the Copyright Office's analysis and with the fact-specific structure of § 107. It could also weaken incentives for lawful acquisition. If two developers use identical works for identical model functions, but one acquires the works through authorized channels while the other knowingly obtains unauthorized copies, the

law need not pretend those facts are identical. The weight assigned to the distinction remains part of the fair-use inquiry.

Documentation is what allows courts to draw such distinctions. A developer that records source, access method, and model use can identify whether a challenged work came from a licensed corpus, an open source, a public-domain collection, or another repository. A developer that cannot reconstruct those facts leaves the court with less reliable evidence. That absence should not be transformed into a doctrinal rule that provenance has no relevance.

C. Poor recordkeeping should not create an evidentiary advantage.

The fair-use inquiry depends on evidence. Rule 56 likewise requires the parties to support factual positions with materials in the record. Fed. R. Civ. P. 56(c). The technical scale of a system does not change those principles.

Consider two developers engaged in comparable conduct. Developer A preserves detailed records identifying source collections, dataset versions, processing steps, and the models trained on each dataset. Developer B preserves little. When a rights-holder identifies a work, Developer A can answer whether the work was used, where it came from, and which model incorporated the relevant dataset. Developer B responds that the corpus was too large or the process too complex to reconstruct.

Copyright law should not make Developer B better situated because it retained less information. Such a rule would invert the incentives created by

modern AI governance. Organizations that invest in documentation would create evidence that can be tested in court. Organizations that do not would gain the benefit of uncertainty. The predictable consequence would be weaker recordkeeping.

CAIDP's concern is not a request for a new adverse-inference rule. Existing procedural and evidentiary law determines the consequences of missing records in particular circumstances. The narrower principle is that courts should not convert an absence of documentation into affirmative support for fair use. A party asserting a factual proposition still must support it with evidence. A party relying on fair use still must address the facts relevant to § 107.

This principle is consistent with AI auditing scholarship. Birhane and her coauthors describe how audits can become performative when reviewers lack access to the information needed to test claims or when institutional structures prevent findings from producing accountability. Birhane et al., supra. The judiciary avoids that problem through ordinary evidentiary rules. It asks what occurred, what evidence supports that account, and whether genuine disputes remain.

The principle is also consistent with contextual data governance. Nissenbaum's work emphasizes that information flows are evaluated by reference to actors, source, purpose, and transmission conditions. Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* 140–47 (Stanford Univ. Press 2010). Copyright is a different body of law, but a similar

factual discipline applies. A court cannot evaluate the legal significance of a data practice if the actors responsible for that practice cannot identify its source or pathway.

The Court can preserve sound incentives by applying ordinary doctrine. It can distinguish among acts. It can consider provenance when the relevant factor permits. It can require factual claims to rest on record evidence. And it can avoid any categorical assumption that technical scale excuses a developer from knowing how its training data were acquired and used.

III. TRAINING-DATA DOCUMENTATION PROTECTS RIGHTS AND SUPPORTS RESPONSIBLE INNOVATION

A. Documentation can coexist with protection for trade secrets and proprietary information.

Transparency does not require indiscriminate public disclosure. AI developers may have legitimate interests in confidential business information. Training pipelines can reveal information about filtering strategies, data weighting, model architecture, security controls, evaluation techniques, and commercial plans.

Those interests are compatible with preservation and targeted disclosure. Courts routinely use protective orders to permit discovery of confidential material. Companies maintain records for legal, financial, safety, and regulatory purposes without publishing all of those records. AI developers can do the same.

A useful distinction is between internal documentation, legally compelled access, and public reporting. Internal records can be granular. Courts and regulators can obtain more detailed information when authorized and can impose

confidentiality protections. Public disclosures can provide categories, source descriptions, and policy information without exposing every proprietary detail.

California AB 2013 follows this structure. It requires specified information about generative-AI training datasets, including source categories and intellectual-property status, but it does not require public disclosure of model weights or source code. CAIDP defended that distinction in *X.AI v. Bonta* because basic training-data information enables oversight while leaving substantial room for protection of legitimate confidential interests.

Mantelero's impact-assessment work likewise shows that accountability frameworks can be structured around necessary information rather than unrestricted disclosure. Mantelero, supra. The purpose of documentation is to permit evaluation of legal and social effects. The amount and recipient of disclosure can be calibrated to the function being served.

B. Documentation supports licensing, auditing, compliance, and dispute resolution.

Training-data records create practical benefits for both developers and rights-holders. First, they support licensing. A developer cannot efficiently license training material if it cannot identify relevant sources, categories, or rights. A rights-holder cannot evaluate a proposed license without enough information to understand the intended use.

The Copyright Office found that voluntary licensing markets for AI training were developing in several sectors. Copyright Office Report at 69–74. Those

markets depend on information. Developers need to know what material they seek to use. Rights-holders need to know how their works will be incorporated. Recordkeeping reduces the transaction costs of both processes.

Second, documentation supports compliance. A dataset may include licensed works, public-domain works, material subject to contractual restrictions, personal information, or content governed by other legal rules. Provenance records permit a developer to apply different rules to different sources and to demonstrate those choices later.

Third, documentation supports auditing. Auditors can compare an organization's stated policies with its actual data pipeline. They can determine whether excluded sources were in fact excluded, whether licenses were followed, and whether a model version used a dataset that should not have been used. Without records, an audit often becomes an assessment of outputs rather than a test of development practices.

Fourth, documentation narrows disputes. A claimant can identify a work. A developer can determine whether the work was present. If the work was never used, the dispute may end quickly. If it was licensed, the developer can produce the license. If it entered through a disputed source, the parties can focus on the legal significance of that source. Accurate records reduce the need for broad inference from model behavior.

Fifth, documentation supports correction. A developer that discovers a problematic source can identify affected datasets and model versions. It can change collection rules, rebuild a dataset, alter future training, or address specific claims. That capacity is part of responsible system management.

The point is not that documentation eliminates copyright disputes. It makes those disputes more precise. It replaces generalized claims about enormous corpora with evidence about particular sources, works, datasets, and uses.

C. Clear accountability rules provide certainty for responsible developers.

The claim that documentation necessarily impedes innovation is incomplete. Documentation has costs. It also creates legal and operational benefits. Bradford's scholarship on digital regulation rejects the assumption that innovation and legal safeguards exist on opposite sides of a fixed tradeoff. Bradford, supra. Rules can shape innovation by rewarding firms that build compliance into product development.

That logic applies to provenance. A developer that knows its data sources can make informed decisions. It can rely on fair use where counsel concludes that fair use applies. It can obtain licenses where licensing is appropriate. It can exclude sources that create unacceptable legal or operational risk. It can respond to rights-holders with evidence rather than uncertainty.

That principle promotes competition as well as compliance. Smaller developers may rely on licensed, curated, or public-domain corpora and maintain

clear records as a way to reduce legal risk. A rule that allows larger developers to invoke scale as a reason they cannot identify their sources would disadvantage firms that invest in disciplined data governance.

Clear rules also improve cross-border operation. AI developers increasingly face documentation duties in multiple jurisdictions. Systems that maintain dataset lineage and model-development records are better positioned to respond to those duties. Documentation therefore forms part of technical and organizational capacity, not merely a litigation burden.

IV. THE COURT SHOULD APPLY EXISTING COPYRIGHT LAW CONSISTENTLY WITH AI ACCOUNTABILITY

A. Human responsibility remains essential when automated systems operate at scale

Artificial intelligence is developed and deployed by people and organizations. Automated collection, filtering, and training do not change that allocation of responsibility. Dignum's work places responsibility on the surrounding sociotechnical system. Russell's work likewise rejects the premise that advanced machine capability relieves humans of control or responsibility. Dignum, supra; Russell, supra.

The Copyright Act operates on the same institutional premise. It regulates acts undertaken by legal persons. A company's use of automated tools to copy or process works does not create a separate category outside the statute. The Court asks what the defendant did, for what purpose, in what amount, and with what market effects.

OpenAI describes extensive organizational control over model development. It identifies training-data construction, processing, pretraining, post-training, evaluation, and deployment as stages in its development process. OpenAI Br. at 5–9. It also identifies which model generations used Library Genesis compilations. Id. at 3. Those descriptions demonstrate that large-scale model development can be documented in model-specific terms.

The Court should therefore reject any implication that the volume of training data makes provenance conceptually unavailable or legally irrelevant. Scale can make documentation difficult. It also makes systematic documentation more important. Modern software systems routinely use logs, version control, access records, dependency records, and data lineage precisely because no individual can remember every event in a complex system.

The Universal Guidelines for Artificial Intelligence state the same principle at a higher level: human responsibility for AI systems must remain identifiable. *Universal Guidelines for AI*, Guideline 6. The Council of Europe Framework Convention likewise adopts accountability and responsibility as governing principles. Those frameworks do not add elements to § 107. They support a cautious judicial approach to arguments that technical complexity should reduce the information expected from sophisticated developers.

B. General-purpose AI developers should preserve sufficient evidence of training-data provenance and use

CAIDP does not ask the Court to prescribe a technical recordkeeping standard. The appropriate architecture will vary across developers and systems. A judicial opinion in a copyright case need not specify database schemas, retention periods, or disclosure formats.

The relevant principle is functional. A developer should preserve enough information to permit meaningful evaluation of legally significant questions concerning training data. Depending on the record, useful information can include principal sources, acquisition methods, collection periods, dataset versions, licenses or known restrictions, filtering and deduplication steps, model-dataset relationships, and the disposition of source or intermediate copies.

These categories are consistent with established machine-learning documentation practices. Datasheets document datasets. Model cards document models. Audit logs document events. Data-lineage systems connect transformations across a pipeline. Impact assessments document risks and responses. No single mechanism is sufficient. Together they show that complex AI development can be accompanied by structured records.

For copyright, the practical use is direct. Records can show whether a work was present, how it entered the corpus, which models used it, and what happened to copies. Those facts allow the Court to apply existing law. They also allow developers to establish defenses with greater precision.

C. The Court should avoid an incentive structure that rewards opacity.

The Court can resolve the pending motions without creating a new regulatory regime. It can apply several propositions that follow from existing copyright and procedural law.

First, fair use applies to identified uses, not to "AI training" as an abstract category.

Second, source and access can be relevant to the character of a use without being dispositive.

Third, factual uncertainty does not become affirmative evidence merely because a system is large. Rule 56 still requires support for factual positions in the record.

Fourth, legitimate confidential information can be protected while permitting sufficient disclosure to adjudicate the claims.

Fifth, a developer should not receive a stronger fair-use position because it kept fewer records of data acquisition and use.

These propositions are modest. They do not decide whether any particular model-training use is fair. They do not establish a general licensing requirement. They do not impose public disclosure of proprietary information. They preserve the factual inquiry that § 107 requires and the evidentiary discipline that Rule 56 requires.

They also align legal incentives with responsible AI development. A developer that documents its sources can demonstrate lawful acquisition where the

facts support that conclusion. It can show that a particular model did not use a disputed dataset. It can establish the scope and purpose of copying. It can address licensing and market questions with evidence. The legal system should encourage those capabilities.

The alternative would create a regulatory inversion. The developer that knows more about its system could face more precise scrutiny, while the developer that knows less could invoke uncertainty as a defense. That outcome would weaken both copyright adjudication and AI accountability.

The Court should instead preserve a straightforward principle: technical scale does not excuse institutional responsibility for evidence. General-purpose AI developers can innovate, rely on fair use where the statute permits, protect trade secrets, and still maintain records sufficient to explain where their training data came from and how those data were used.

CONCLUSION

The Court should apply § 107 to the specific acts and record before it. It should not adopt a categorical rule that all generative-AI training is fair use or that all generative-AI training is infringing. Acquisition, dataset construction, model training, retention, transfer, and outputs can present distinct questions and should be evaluated separately when the claims and evidence require that distinction.

The Court should also recognize the evidentiary role of training-data provenance. Source and lawful access can bear on the character of a use.

Documentation permits courts to determine what occurred, permits developers to demonstrate lawful practices, permits rights-holders to evaluate claims, and supports licensing, auditing, compliance, and efficient dispute resolution.

Technological scale does not eliminate responsibility. Inadequate recordkeeping should not create a legal advantage. The Court should apply existing copyright law in a manner that preserves incentives for developers of general-purpose AI systems to know, document, and account for the data on which their systems depend.

Respectfully submitted,

CENTER FOR AI AND DIGITAL POLICY

/s/ Marc Rotenberg
Marc Rotenberg
Counsel for Amici Curiae
Center for AI and Digital Policy, AI Experts, and Legal Scholars
1100 13th St. NW, Suite 800
Washington, D.C. 20005
+1 202 415 6788
rotenberg@caidp.org

Dated: October 30, 2026

CERTIFICATE OF COMPLIANCE

I certify that this brief complies with the Court's August 7, 2026 order because the brief contains 5,926 words, excluding the caption, tables, signature block, and this certificate. The brief uses Times New Roman 14-point type.

/s/ Marc Rotenberg
Marc Rotenberg