United States District Court
Southern District of New York

In re OpenAI, Inc., Copyright Infringement Litigation This document relates to: All Actions	**No. 25-md-3143 (SHS) (OTW)**

Proposed brief of the Foundation for American Innovation, David R. Shedd, Lieutenant General Michael Groen, USMC (Ret.), and George Barnes as amici curiae in support of OpenAI on the fourth fair-use factor

Tim Hwang

Foundation for American Innovation

408 C St NE, Washington, DC 20002

973-960-4955

tim.hwang@thefai.org

Counsel for proposed amici curiae

Pro hac vice admission pending

Corporate, authorship, and funding disclosure

The Foundation for American Innovation is a nonprofit corporation described in § 501(c)(3) of the Internal Revenue Code. It has no parent corporation and issues no stock; no publicly held corporation owns any part of it.

David R. Shedd, Lieutenant General Michael Groen, U.S. Marine Corps (Ret.), and George Barnes appear solely in their individual capacities. Titles and affiliations are provided for identification only.

No party's counsel authored this brief in whole or in part. No party or party's counsel contributed money intended to fund its preparation or submission. No person other than amici or their counsel contributed money intended to fund its preparation or submission.

Table of contents

Corporate, authorship, and funding disclosure i

Interests of amici 1

Issue presented 3

Introduction 3

Argument 4

I. General model capabilities serve concrete national-security missions. 4

A. Congress has connected AI training and data to defense capability. 4

B. The Government obtains general capabilities through commercial frontier models. 5

C. Examples show how agencies apply general capabilities to national-security missions. 6

D. Strategic competition magnifies the cost of lost capability. 8

II. *Oracle* supports considering the public benefits of capabilities learned through training. 9

A. *Oracle* places pertinent public benefits within factor four. 9

B. Second Circuit law keeps these public benefits distinct from copyright-market harm. 10

C. These benefits are downstream of corpus-scale training that produces reusable general capabilities....12

III. The consequences of restricting training data belong in the public-benefit analysis. 13

A. A permission requirement may restrict access to useful training data.....13

B. Losses in model capabilities can affect national-security applications.....14

C. The Court should weigh these public benefits under factor four. 15

Conclusion 15

Appendix — List of Amici Curiae....17

Table of authorities

Cases

Am. Geophysical Union v. Texaco Inc., 60 F.3d 913 (2d Cir. 1995).....11

Am. Soc'y for Testing & Materials v. UpCodes, Inc., 172 F.4th 253 (3d Cir. 2026) 11

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 11 F.4th 26 (2d Cir. 2021).....11

Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) 10–11

Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) 10–11

Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605 (2d Cir. 2006) 10

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) 10

Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018).....11

Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021).....3–4, 9–10, 15

Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163 (2d Cir. 2024) 10–11

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

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

MCA, Inc. v. Wilson, 677 F.2d 180 (2d Cir. 1981) 10

Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984) 3

Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P., 756 F.3d 73 (2d Cir. 2014).....10

Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc., 765 F. Supp. 3d 382 (D. Del. 2025), aff'd, No. 25-2153 (3d Cir. Sept. 29, 2026) 11

Statutes

17 U.S.C. § 107.....9

17 U.S.C. §§ 106, 502–504.....14

Servicemember Quality of Life Improvement and National Defense Authorization Act for Fiscal Year 2025, Pub. L. No. 118-159, 138 Stat. 1773 (2024) 4–5

National Defense Authorization Act for Fiscal Year 2026, Pub. L. No. 119-60, 139 Stat. 718 (2025) 5

Other authorities

Anthropic, *Assessing Claude Mythos Preview's Cybersecurity Capabilities* (2026) 7–8

Autorité de la concurrence, *Opinion 24-A-05* (2024) 14

Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models* (2021, rev. 2022) 6, 14

Nicholas Carlini et al., *Extracting Training Data from Large Language Models* (2021) 12

Declaration of Chris Callison-Burch, ECF No. 1746 (2026) 12

Defense Advanced Research Projects Agency, *AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense* (2025) 7

Department of War, *Artificial Intelligence Strategy for the Department of War* (2026) 6

Department of War, *Classified Networks AI Agreements* (2026) 6

Charles Duhigg & Keith Bradsher, *How the U.S. Lost Out on iPhone Work* (2012) 12

Foundation for American Innovation, *About Us* 1

Foundation for American Innovation, Amicus briefs in Ross, Bartz, and Anthropic 1

Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* (2022) 13

Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (2020) 6, 14

Shayne Longpre et al., *A Pretrainer's Guide to Training Data* (2024) 4, 13

National Development and Reform Commission et al., *Data-Labeling Industry Implementation Opinions* (2024) 8

National Institute of Standards and Technology, *Adversarial Machine Learning* (2025) 14

National Institute of Standards and Technology, *Foundation Model*.....3, 12

National Security Commission on Artificial Intelligence, *Final Report* (2021) 8

National Security Presidential Memorandum/NSPM-11, *Artificial Intelligence in the National Security Enterprise* (2026).....5–6, 8

New York Times Co. v. Microsoft Corp., Complaint, No. 23-cv-11195, ECF No. 1 (S.D.N.Y. Dec. 27, 2023).....12

Office of the Director of National Intelligence & Central Intelligence Agency, *The INT of First Resort* (2024).....7, 12

State Council, *Opinions on Deepening Implementation of the "AI Plus" Initiative* (2025) 8

Statement of Interest of the United States, *In re OpenAI, Inc., Copyright Infringement Litigation* (2026) 2, 6

Statement of Vice Adm. Frank D. Whitworth Before the House Armed Services Committee (2025) 7

U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3* (2025).....14

Interests of amici

The Foundation for American Innovation is a nonprofit public-policy organization. Its work spans artificial intelligence, technology and statecraft, and governance.[1]

FAI has studied copyright, AI training, government access to advanced models, and technological competition. It has also filed amicus briefs on those subjects in the Third, Ninth, and D.C. Circuits.[2]

David R. Shedd served as Deputy Director and later Acting Director of the Defense Intelligence Agency. He previously served as Deputy Director of National Intelligence for Policy, Plans, and Requirements, overseeing the development and implementation of Intelligence Community policies and strategic planning. He also served as the NSC's Senior Director and Special Assistant to the President for Intelligence.

Lieutenant General Michael Groen, U.S. Marine Corps (Ret.), served as Director of the Department of Defense's Joint Artificial Intelligence Center, where he led efforts to integrate artificial intelligence into military capabilities and departmental processes. He previously served as Director for Intelligence on the Joint Staff and in a senior computer-network-operations role at the National Security Agency.

---

[1] *About Us*, Found. for Am. Innovation, https://www.thefai.org/about-us (last visited Oct. 25, 2026).

[2] Corrected Brief for Amicus Curiae Foundation for American Innovation in Support of Defendant-Appellant, *Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc.*, No. 25-2153, ECF No. 62 (3d Cir. Sept. 30, 2025); Brief of Foundation for American Innovation as Amicus Curiae Supporting Petitioners and Rule 23(f) Leave to Appeal, *Bartz v. Anthropic PBC*, No. 25-4843, Dkt. Entry No. 10 (9th Cir. Aug. 7, 2025); Brief of Amici Curiae Foundation for American Innovation et al. in Support of Petitioner's Emergency Motion for a Stay, *Anthropic, PBC v. U.S. Dep't of War*, No. 26-1049 (D.C. Cir. Mar. 11, 2026), https://cdn.sanity.io/files/d8lrla4f/staging/296ca8b253f08fd869074c8dce015bda96345a5c.pdf (last visited Oct. 25, 2026).

George Barnes served as the Deputy Director of the U.S. National Security Agency. He currently engages on national security with educational institutions and non-profit organizations, builds and scales new companies in the venture ecosystem, and advises companies in national security and emerging technologies.

Together, FAI's policy work and the individual amici's experience in intelligence, defense, and military AI inform amici's shared interest in the national-security benefits of capabilities developed through AI training.

David R. Shedd, Lieutenant General Michael Groen, U.S. Marine Corps (Ret.), and George Barnes appear solely in their individual capacities. Titles and affiliations are provided for identification only.

The United States has addressed AI's public benefits and national-security importance. Amici supplement that submission by explaining how training data supports particular national-security capabilities.[3] OpenAI and the authors and publishers address the disputed facts about particular works, models, licenses, and outputs. Amici address how factor four should consider a public benefit created through corpus-scale training: the general capabilities of the resulting model and the national-security systems that use them.

[3] *Statement of Interest of the United States* at 2, 8, 17–19, *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 25-md-3143, ECF No. 1682 (S.D.N.Y. Sept. 1, 2026), https://chatgptiseatingtheworld.com/wp-content/uploads/2026/09/Statement-of-Interest-of-the-United-States-on-Fair-Use-Sept-1-2026-gov.uscourts.nysd_.640396.1682.0.pdf.

Issue presented

Foundation models learn general capabilities through self-supervised training across broad collections of data, or corpora. The Government now uses those capabilities for defense, intelligence, and cybersecurity missions. Does factor four count the resulting public benefits when it weighs the challenged training use?[4]

Introduction

General AI capabilities already serve important national security purposes. Over the last several years, the Government has acquired and adapted commercial frontier models for intelligence analysis, cyber defense, operational planning, and decision support. Both branches of government understand the importance of this new technological paradigm: Congress has supported it by tying defense AI to training and data and the executive by proclaiming rapid access to current models as one of its top national security priorities.

Copyright protects authors for a simple reason: their work advances knowledge and culture. Fair use serves the same end. It preserves room for new uses that benefit the public without displacing the expression copyright protects.[5] But factor four of the fair use test is not limited to an accounting of the copyright holder's potential lost revenue. Under *Google LLC v. Oracle Am., Inc.,* courts may also consider the public benefits that copying enables, particularly where the use puts copyrighted works toward new productive ends. To the extent AI training enables the creation of new knowledge, expression, software, and other capabilities, those

---

[4] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 35–36 (2021); *Foundation Model*, Nat'l Inst. of Standards & Tech., Computer Security Resource Center Glossary, https://csrc.nist.gov/glossary/term/foundation_model (last visited Oct. 25, 2026); *see infra Part I.*

[5] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 16–18, 35–36 (2021); *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 429 (1984); *Mazer v. Stein*, 347 U.S. 201, 219 (1954).

benefits bear on factor four's assessment of the use's market effects. That reasoning supports a novel application here: considering the national-security value of capabilities that enable new analysis, software, and other productive work.

The link between copying a newspaper article and developing national-security capabilities lies in the training process. During pretraining, models identify shared patterns across a broad corpus of information. Learning those patterns helps develop language, synthesis, translation, coding, and reasoning capabilities.

Requiring permission for the challenged training could impair those benefits if it materially restricts access to useful training data. Removing useful sources from training data can reduce capabilities used by downstream systems, which could include the national security systems the Government acquires and adapts. Under *Oracle*, those public benefits belong in the factor-four balance alongside cognizable market harm.[6]

Argument

I. General model capabilities serve concrete national-security missions.

General AI capabilities serve important national-security functions, from intelligence to cyber defense to operational planning to decision support. Legislation, procurement policy, and adoption show the Government's assessment of their importance.

A. Congress has connected AI training and data to defense capability.

In the fiscal year 2025 defense act, Congress directed the Department of Defense to expand computing resources for training, fine-tuning, modifying, and deploying large AI

---

[6] *See Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 35–36 (2021); Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) 3245*, 3251–52 (2024), https://aclanthology.org/2024.naacl-long.179/.

systems. Separately, Congress directed the Department to develop general-purpose AI systems trained using Department data or data curated from "general, openly or commercially available sources." Congress also required budget plans for the training data needed to train, maintain, and improve defense AI, including acquisition and sustainment costs.[7] The following year, Congress authorized National Security and Defense Artificial Intelligence Institutes focused on "cross-cutting challenge[s] or foundational science" and expressly authorized them to make public- and private-sector datasets available for "training and testing artificial intelligence systems" for national security and defense.[8]

Congress has thus consistently treated AI training data as an integral part of the creation and deployment of military capabilities. These enactments establish the importance of the asserted public benefit.

B. The Government obtains general capabilities through commercial frontier models.

The executive has likewise made access to commercial frontier models a national-security priority. The President's June 2026 national-security directive, NSPM-11, states that AI will be "among the most transformative technologies to national security in the history of the United States," directing that frontier models be made available to national-security personnel

[7] Servicemember Quality of Life Improvement and National Defense Authorization Act for Fiscal Year 2025, Pub. L. No. 118-159, §§ 1532(b)(1), (b)(2)(A)(i), (d)(1)–(2), 1533(a)–(b), 138 Stat. 1773, 2144–46 (2024), https://www.congress.gov/118/plaws/publ159/PLAW-118publ159.pdf.

[8] National Defense Authorization Act for Fiscal Year 2026, Pub. L. No. 119-60, § 224(a), (b)(1), (c)(2)(A), 139 Stat. 718, 782–83 (2025), https://www.govinfo.gov/content/pkg/PLAW-119publ60/pdf/PLAW-119publ60.pdf.

without delay.[9] The Department's January 2026 AI Strategy makes the ability to deploy a vendor's newest model within 30 days of public release a primary procurement criterion.[10]

Reported use of the Department's GenAI.mil platform shows substantial adoption. In May 2026, the Department reported more than 1.3 million users of GenAI.mil collectively engaging in tens of millions of prompts and deploying hundreds of thousands of AI agents. The Department also signed agreements with OpenAI and seven other providers to run models on the Department's classified networks.[11]

These applications draw on capabilities developed during the commercial pretraining at issue in this case. It is true that mission-specific prompting, retrieval, tools, and fine-tuning can supplement those capabilities, but improvements or losses in the underlying model can also impact the quality of the tools available to national-security users.[12]

C. Examples show how agencies apply general capabilities to national-security missions.

Three examples show how agencies use or seek to use these capabilities.

---

[9] National Security Presidential Memorandum/NSPM-11, *Artificial Intelligence in the National Security Enterprise* §§ 1, 2(a)–(b), 4(a) (June 5, 2026), https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/. *See* also *Statement of Interest of the United States*, supra note 3, at 2, ECF No. 1682.

[10] Department of War, *Artificial Intelligence Strategy for the Department of War: Accelerating America's Military AI Dominance* 1, 4 (Jan. 9, 2026), https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.

[11] Department of War, *Classified Networks AI Agreements* (May 1, 2026), https://www.war.gov/News/Releases/Release/Article/4475177/classified-networks-ai-agreements/.

[12] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models 3–7* (Ctr. for Rsch. on Found. Models, Stanford Inst. for Hum.-Centered Artificial Intelligence 2021, rev. July 12, 2022), https://doi.org/10.48550/arXiv.2108.07258; Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 33 Advances in Neural Information Processing Systems 9459, 9459–62 (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html; National Security Presidential Memorandum/NSPM-11, *Artificial Intelligence in the National Security Enterprise* §§ 2(a)–(b), 4(a) (June 5, 2026), https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/.

First, open-source intelligence. Open-source intelligence is the tradecraft of learning from what the world publishes—information that changes at "breakneck speed." The Intelligence Community calls it "the INT of first resort." Its 2024 strategy explains that generative AI can help identify themes and patterns and summarize large volumes of text.[13]

Second, geospatial intelligence. Geospatial intelligence includes analysis of satellite and other overhead imagery. The National Geospatial-Intelligence Agency has described "geospatial agents" that enable large language models to access geospatial data. Separately, it has described pilots of generative AI models "to sift through millions of detections and GEOINT reports," improving the speed and accuracy of target identification and reducing delays in disseminating intelligence.[14]

Third, cyber defense. Cyber defense uses general coding and reasoning capabilities: a model that can read and write software may help find the flaws an attacker would exploit—and develop patches. DARPA's AI Cyber Challenge, a two-year competition to find and fix flaws in the open-source software that underlies critical infrastructure, gave competing teams model and cloud support from OpenAI, Anthropic, Google, and Microsoft. In the 2025 final, the systems, which combined language models with other cybersecurity tools, analyzed more than 54 million lines of code, found 54 synthetic and 18 real vulnerabilities, and submitted patches in an average of 45 minutes.[15] Anthropic likewise attributes Mythos Preview's advanced vulnerability-

---

[13] Office of the Director of National Intelligence & Central Intelligence Agency, *The INT of First Resort: Unlocking the Value of OSINT—The IC OSINT Strategy 2024–2026*, at 1–5 (Mar. 2024), https://www.dni.gov/files/ODNI/documents/IC_OSINT_Strategy.pdf.

[14] Statement of Vice Adm. Frank D. Whitworth, Director, National Geospatial-Intelligence Agency, Before the Subcomm. on Strategic Forces of the H. Armed Servs. Comm. 5 (May 14, 2025), https://www.nga.mil/assets/files/FY26_VADM_Whitworth_Testimony_HASC_National_Security_Space_Programs.pdf.

[15] Defense Advanced Research Projects Agency, *AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense* (Aug. 8, 2025), https://www.darpa.mil/news/2025/aixcc-results.

discovery and exploitation capabilities to general improvements in coding, reasoning, and autonomy.[16]

D. Strategic competition magnifies the cost of lost capability.

These missions are part of a contest for technological advantage. The congressionally chartered National Security Commission on Artificial Intelligence warned in its 2021 report that AI will shape future military advantage and that, in this field, China aims to overtake the United States. Current policy treats continued American AI leadership as a national-security objective.[17]

China's policies likewise treat data supply as a foundation for AI capability. Its 2025 "AI Plus" policy calls for high-quality AI datasets and lawful access to publicly funded copyrighted content, and encourages exploring compensation incentives for data suppliers. A separate policy sets a goal of more than 20% compound annual growth in the data-labeling industry by 2027 and calls for industry datasets to support AI applications. These measures underscore that access to useful data is itself a dimension of AI competition.[18]

A demonstrated loss of American AI capability may carry a strategic cost as competitors develop their own systems.

---

[16] Anthropic, *Assessing Claude Mythos Preview's Cybersecurity Capabilities* (Apr. 7, 2026), https://www.anthropic.com/research/mythos-preview.

[17] National Security Commission on Artificial Intelligence, *Final Report* 2, 7, 61, 160–61 (2021), https://www.govinfo.gov/content/pkg/GOVPUB-Y3-PURL-gpo153246/pdf/GOVPUB-Y3-PURL-gpo153246.pdf; National Security Presidential Memorandum/NSPM-11, *Artificial Intelligence in the National Security Enterprise* §§ 1–2 (June 5, 2026), https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/.

[18] State Council, *Opinions on Deepening Implementation of the "AI Plus" Initiative*, Part III(8), Guo Fa [2025] No. 11 (Aug. 21, 2025), https://www.cac.gov.cn/2025-08/27/c_1758018277755538.htm; National Development and Reform Commission et al., *Implementation Opinions on Promoting High-Quality Development of the Data-Labeling Industry*, Parts I, II(2), Fa Gai Shu Ju [2024] No. 1822 (Dec. 26, 2024; published Jan. 13, 2025), https://www.nda.gov.cn/sjj/zwgk/zcfb/0113/20250113101518513076594_mobile.html (titles translated from Chinese).

II. *Oracle* supports considering the public benefits of capabilities learned through training.

The national-security missions detailed above yield public benefits relevant to factor four under *Oracle*.

A. Oracle places pertinent public benefits within factor four.

Factor four begins with the use's effect on the potential market for or value of the copyrighted work.[19] Market harm remains central to this analysis. But the Supreme Court has rejected the premise that potential revenue loss ends the inquiry. In *Oracle*, the Court considered commercial copying of part of a software interface and explained that it "must take into account the public benefits the copying will likely produce." In that case, the effect of Google's copying in allowing programmers to "make use of their knowledge and experience" using the Sun Java API when writing new smartphone programs supported the Court's fair-use analysis.[20]

Oracle also identifies the kind of benefit relevant here. In that case, the Court did not limit cognizable benefits to particular works produced downstream. Rather, it credited Google's creation of a new technological platform that allowed programmers to "put their accrued talents to work," treating that enabling capability as consistent with copyright's objective of creative progress.[21] Thus, if corpus-scale training produces reusable capabilities that enable new productive activity, those capabilities themselves are relevant public benefits under *Oracle*'s reasoning.

Oracle concerned software interfaces, not AI training, and cautioned that these public-benefit questions are not always relevant. Its reasoning nevertheless applies to training that

---

[19] 17 U.S.C. § 107(4).

[20] *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 34–36, 39–40 (2021).

[21] *Id.* at 30–31, 40.

enables new analysis and software—the "creative production of new expression" copyright promotes.[22] General-purpose AI brings that productive capacity to national-security work, including intelligence analysis and cyber defense.

B. Second Circuit law keeps these public benefits distinct from copyright-market harm.

Recognizing a public benefit is distinct from deciding whether an asserted market harm is one copyright recognizes. The availability of a license does not itself establish that permission is required.[23] Factor four asks both what benefits the use produces and whether the countervailing loss arises from substitution for interests that copyright actually protects.

MCA, Inc. v. Wilson, which *Oracle* cited, described factor four as a comparison between the public benefit of permitting a use and the copyright owner's "expectation of gain" from prohibiting it. *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.* likewise recognized the public value of timely financial information while separately asking whether the use impaired the copyright owner's legitimate interests in protected expression.[24]

The Second Circuit's search cases enforce the same boundary. In addressing full-text search, *HathiTrust* rejected lost licensing revenue because search did not substitute for the books. *Google Books* likewise protected full-corpus copying used to enable search where the resulting service did not provide a significant substitute for protected expression. And *Hachette* rejected fair use where Internet Archive supplied complete books as substitutes for authorized copies.

---

[22] *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 30–31, 35–36, 40 (2021). *Cf. Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1054–55, 1060 (N.D. Cal. 2025) (treating market dilution from noninfringing outputs as potentially cognizable harm, but contemplating that nonprofit training for purposes such as "national security" "might be fair use even in the face of some amount of market dilution").

[23] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 585 n.18 (1994) ("If the use is otherwise fair, then no permission need be sought or granted."); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (2d Cir. 2006) (rejecting claimed licensing harm from the transformative use at issue).

[24] *MCA, Inc. v. Wilson*, 677 F.2d 180, 183 (2d Cir. 1981); *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73, 90–92 (2d Cir. 2014).

Even there, the court weighed "the benefit the public will derive" against the copyright owner's gain, concluding that the asserted benefits did not outweigh the harm to publishers' markets and incentives to create. *TVEyes* likewise shows that a useful analytical service may cause cognizable licensing harm when it redistributes protected broadcasts. *Texaco* and *Ross* assessed the purpose and competitive effects of uses they found nontransformative: archival copying of research articles and use of headnotes to build a competing legal-research tool. In affirming *Ross*, the Third Circuit recognized a potential training-license market and rejected national-security benefits unsupported by evidence linking that platform to those uses, while distinguishing generative AI. Neither makes an available license dispositive.[25]

The Third Circuit recently considered both public benefit and market harm in *Am. Soc'y for Testing & Materials v. UpCodes, Inc.* There, *UpCodes* made freely available copyrighted technical standards that had been incorporated by reference into binding law. The court placed the "clear and significant public benefit" of enhanced public access to law within factor four, while separately assessing the cognizable economic harm from substitution and the possibility that lost revenue could impair ASTM's incentives and ability to develop new standards. The court found factor four equivocal on that preliminary-injunction record.[26]

---

[25] *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 99–100 (2d Cir. 2014); *Authors Guild v. Google, Inc.*, 804 F.3d 202, 214–18, 223–25 (2d Cir. 2015); *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 194–96 (2d Cir. 2024) (quoting *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 11 F.4th 26, 48 (2d Cir. 2021)); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 179–81 (2d Cir. 2018). *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 922–23, 929–31 (2d Cir. 1995); *Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc.*, 765 F. Supp. 3d 382, 398–401 (D. Del. 2025), aff'd, No. 25-2153, slip op. at 17 n.7, 26–27 (3d Cir. Sept. 29, 2026).

[26] *Am. Soc'y for Testing & Materials v. UpCodes, Inc.*, 172 F.4th 253, 270–71 (3d Cir. 2026), https://www2.ca3.uscourts.gov/opinarch/242965p.pdf.

C. These benefits are downstream of corpus-scale training that produces reusable general capabilities.

Foundation models are trained to acquire capabilities that can be reused across many downstream tasks. NIST defines a foundation model as one trained on broad data through self-supervised learning and adaptable to many downstream tasks. During next-token pretraining, models predict the next unit of text, such as a word or part of a word, and adjust their numerical weights, or parameters, to improve those predictions. Through this process, the model learns statistical relationships distributed across the corpus. These general capabilities need not correspond to any single source, although models can also memorize training material.[27]

That shared structure explains why corpus breadth matters to the development of general capabilities. Varied sources expose the model to different languages, domains, formats, and relationships.

Consider, as an illustration, the Times article "*How the U.S. Lost Out on iPhone Work*," identified in its complaint. Training on reporting of this kind exposes a model to descriptions of suppliers, production constraints, and corporate decisions. Across many works, recurring patterns can contribute to general capabilities for comparing and synthesizing reports. An analyst could apply those capabilities to new reporting about supply-chain vulnerabilities.[28]

---

[27] *Foundation Model*, supra note 4; Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, in 30th USENIX Security Symposium 2633, 2633–35 (2021), https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. *See also* Declaration of Chris Callison-Burch ¶¶ 72–77, ECF No. 1746.

[28] Charles Duhigg & Keith Bradsher, *How the U.S. Lost Out on iPhone Work*, N.Y. Times (Jan. 21, 2012), identified in Complaint ¶ 100 & n.27, *New York Times Co. v. Microsoft Corp.*, No. 23-cv-11195, ECF No. 1 (S.D.N.Y. Dec. 27, 2023), https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf; Office of the Director of National Intelligence & Central Intelligence Agency, The INT of First Resort: Unlocking the Value of OSINT—The IC OSINT Strategy 2024–2026, at 4–5 (Mar. 2024), https://www.dni.gov/files/ODNI/documents/IC_OSINT_Strategy.pdf.

Empirical research shows that training-data scale and composition affect model performance. In one study, research scientists trained more than 400 models on 5 billion to 500 billion tokens. Their results indicated that compute-optimal performance required the number of training tokens to be scaled equally with model size. At the same training-compute budget, their 70-billion-parameter Chinchilla model, trained on 1.4 trillion tokens, outperformed the 280-billion-parameter Gopher model, trained on 300 billion tokens. On the RACE-h reading-comprehension benchmark, the models scored 82.3% and 71.6%, respectively.[29] Separate experiments found that removing broad web and book sources from pretraining reduced average performance across question-answering tasks, even after task-specific fine-tuning.[30]

The capabilities described here are learned across a corpus. No single work need independently create translation, coding, synthesis, or reasoning capability for corpus-scale training to produce those capabilities collectively. Restrictions that impair those capabilities may therefore diminish the national-security benefits relevant to the factor-four analysis.

III. The consequences of restricting training data belong in the public-benefit analysis.

A. A permission requirement may restrict access to useful training data.

A ruling confined to substitutive outputs need not restrict training. If the challenged training requires permission, developers may need to license or exclude affected works. Existing

---

[29] Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models*, 35 Advances in Neural Information Processing Systems 30,016, 30,016–17, 30,022–24 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf. The comparison also involved changes to the tokenizer, optimizer, and data mixture.

[30] Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) 3245*, 3250–52, §§ 5–6 & fig. 5 (2024), https://aclanthology.org/2024.naacl-long.179/. The experiments varied both the amount and composition of training data; some filtering choices improved performance.

models could also be affected, depending on the relief ordered.[31] Licensing may be practical for organized catalogs but may be difficult for diffuse sources whose rightsholders are hard to identify. The Copyright Office itself has recognized this limit, stating that a growing licensing market "does not itself establish that voluntary licensing is feasible at scale for all AI training needs." Even as the court in *Kadrey* rejected predictions that licensing would halt AI development, it acknowledged that developers might "move somewhat more slowly." These delays could impair capabilities used for national security. As the studies in Part II.C show, restrictions that materially reduce useful training data can degrade model performance.[32]

B. Losses in model capabilities can affect national-security applications.

That loss can extend to downstream applications. As shown above, a foundation model supplies capabilities used by applications built on it. Prompting, retrieval, and fine-tuning can supplement those capabilities and compensate for some limitations.[33]

The sources used to train a model can affect downstream security. NIST identifies training-data poisoning as a supply-chain risk and explains that malicious backdoors can persist after fine-tuning. A compromised foundation model can, in turn, affect applications built on it. The extent of that risk depends on the sources developers use and the safeguards they adopt.[34]

---

[31] 17 U.S.C. §§ 106(1), 502–504.

[32] U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* 70–71 & n.401, 103–04 (Pre-Publication Version, May 2025) (discussing licensing obstacles and potential collective-licensing mechanisms), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf; Autorité de la concurrence, *Opinion 24-A-05 of 28 June 2024 on the Competitive Functioning of the Generative Artificial Intelligence Sector* ¶¶ 355–61, at 90–91 (June 28, 2024), https://www.autoritedelaconcurrence.fr/sites/default/files/2025-04/24a05_eng_1.pdf. *Kadrey*, 788 F. Supp. 3d at 1059.

[33] *See* Bommasani et al., supra note 12, at 3–4, 85, 88; Lewis et al., supra note 12, at 9459–62.

[34] Apostol Vassilev et al., Nat'l Inst. of Standards & Tech., NIST AI 100-2e2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* 37, 41–43 (Mar. 2025), https://doi.org/10.6028/NIST.AI.100-2e2025; Bommasani et al., supra note 12, at 105–06.

C. The Court should weigh these public benefits under factor four.

Oracle identifies the relevant consideration: "the public benefits the copying will likely produce."[35] The causal question is whether the challenged training produced capabilities the Government uses. If OpenAI's record establishes that link, the Court should weigh the resulting national-security benefits alongside cognizable market harm under factor four. The Court should also consider whether a permission requirement would impair or delay those capabilities.

Conclusion

Amici respectfully request that, under *Oracle*, the Court weigh under factor four the national-security benefits that the record links to the challenged training, alongside cognizable market harm.

[35] *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 35–36 (2021).

Respectfully submitted,

/s/ Tim Hwang

Tim Hwang

Foundation for American Innovation

408 C St NE, Washington, DC 20002

973-960-4955

tim.hwang@thefai.org

Counsel for proposed amici curiae

Pro hac vice admission pending

Dated: October 30, 2026

Appendix — List of Amici Curiae

FOUNDATION FOR AMERICAN INNOVATION

Nonprofit public-policy organization

DAVID R. SHEDD

Former Deputy Director and Acting Director

DEFENSE INTELLIGENCE AGENCY

LIEUTENANT GENERAL MICHAEL GROEN, USMC (RET.)

Former Director

JOINT ARTIFICIAL INTELLIGENCE CENTER

U.S. DEPARTMENT OF DEFENSE

GEORGE BARNES

Former Deputy Director

U.S. NATIONAL SECURITY AGENCY

David R. Shedd, Lieutenant General Michael Groen, U.S. Marine Corps (Ret.), and George Barnes appear solely in their individual capacities. Titles and affiliations are provided for identification only.

Certificate of compliance

This proposed brief contains 4,393 words and complies with Local Civil Rule 7.1(c) and the Court's August 7, 2026 Order, ECF No. 1663, limiting proposed amicus briefs to 6,000 words. The count was determined using the word-processing software used to prepare the document and includes footnotes and the appendix, and excludes the caption, table of contents, table of authorities, signature block, and this certificate. The brief complies with Local Civil Rule 7.1(b) by using 12-point text, 10-point footnotes, one-inch margins, and double-spaced body text.

/s/ Tim Hwang

Tim Hwang