Statement of
BHAMATI VISWANATHAN
Assistant Professor, New England Law | Boston

before the

SENATE COMMITTEE OF THE JUDICIARY
SUBCOMMITTEE ON CRIME AND COUNTERTERRORISM

July 16, 2025

Bhamati Viswanathan, Assistant Professor of Law at New England Law | Boston, submits this statement for the record concerning the hearing titled Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training before the Senate Judiciary Committee, Subcommittee on Crime and Counterterrorism, on July 16, 2025.

The Powerful and Robust U.S. Creative Economy Is Being Irreparably Harmed by the Proliferation of Digital Piracy Undertaken By So-Called Shadow Libraries

The arts and cultural economic activity in the US, as estimated by the U.S. Bureau of Economic Analysis accounted for 4.2 percent of GDP, or $1.17 trillion, in 2023. It is dependent on strong copyright protection.[1] In Article 1 Section 8 Clause 8, the U.S. Constitution establishes this right, creating an incentive structure that creators rely on. Innovation is the goal of copyright, as set forth in the Constitution: "to promote progress in Science and the Useful Arts."

Piracy circumvents that balance. Piracy is the consumption of unlicensed copyrighted products, differing from counterfeiting, which is the consumption of unlicensed trademarked products. Digital piracy, specifically, mirrors supply chain for physical pirated goods in that intermediaries facilitated discovery of pirated contents by consumers. Here, the distribution of content form providers to consumers, and the flow of payments from consumers to both platforms and providers. However, differing from physical piracy, digital piracy does not require manufacturing steps and is distributed virtually, reducing cost and increasing scope and scale of digital piracy operations.[2]

These shadow libraries play significant roles as illicit actors and continue to be pursued by the Federal Bureau of Investigations (FBI) and the Department of Homeland Security (DHS). For example, in 2022, the FBI seized domains associated with Z-Library and charged two of its operators with criminal copyright infringement, wire fraud, and money laundering.[3] Likewise, in its Operations Intangibles, U.S. Immigration and Customs Enforcement's (ICE) of the DHS have also outlined its commitment to "stop digital piracy and eliminate a vital source of illicit revenue from transnational criminal organizations," citing that these activities have continued to feed a

[1] Arts and Cultural Production Satellite Account, U.S. and States, 2023 | U.S. Bureau of Economic Analysis (BEA)
[2] Brett Danaher, Michael D. Smith, and Rahul Telang, Piracy Landscape Study: Analysis of Existing and Emerging Research Relevant to Intellectual Property Rights (IPR) Enforcement of Commercial-Scale Piracy , https://www.cmu.edu/entertainment-analytics/documents/uspto.pdf
[3] Federal Law Enforcement Arrests and Indicts Z-Library Operators with AG's Assistance - The Authors Guild

criminal enterprise whose profits are used to support other organized criminal endeavors, including violent crime and trafficking.[4]

Generative AI Companies Are Relying on Pirated Materials to Build Their Large Language Models and Thereby Augmenting the Harms That Shadow Libraries Cause

GenAI companies are required to ingest vast amounts of materials in order to build robust large language models (LLMs). This is because LLMs are essentially sophisticated prediction models: they learn structure, syntax, speech patterns, and other linguistic foundations, and then "predict" language sequences based on their acquired learning. GenAI companies must find these vast amounts of materials from available digital sources, databases, and repositories.

It is clear that GenAI companies ingest works from pirate sources.[5] When LLM models are trained on pirated works, they circumvent copyright law by training on works that have already been illicitly reproduced, digitized, and distributed. Evidence shows that GenAI companies have willfully, knowingly and repeatedly trained on pirated materials, despite being aware that their source shadow libraries are circumventing the law to obtain and share those materials.[6]

This is a crime compounding a crime: the initial crime is the illicit copying, making available, and distribution of materials under copyright; and the compounded crime is the ingestion and use of these materials in the creation and development of LLMs by GenAI companies.

When Generative AI Companies Ingest Pirated Materials, They Directly Harm Copyright Holders by Undermining Their Rights and Usurping Their Rewards

The training of LLMs on pirated materials is far from a "victimless" crime. Authors, artists, filmmakers, and photographers are among the creators whose works are taken and used without permission or payment.[7] Publishers, film producers and distributors, newspapers, and media outlets are among the intermediaries whose commercial services are usurped, also without permission or payment.[8] These are real victims: they relied on well-established copyright laws to protect their original works,[9] only to have those works taken *en masse* to build LLMs that in turn can enable the mass production of infringing works. And this harm is more than hypothetical: there is a direct relationship between the rise of e-book piracy and the decline in authors'

[4] Operation Intangibles | ICE

[5] *E.g.*, *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC, Dkt. No. 567-45 (Meta email noting "GenAI has been approved to use LibGen for Llama 3" despite acknowledging that LibGen is "a dataset we know to be pirated"); Dkt. No. 567-25 (Meta employee stating that "It's the piracy (and us knowing and being accomplices) that's the issue."); Dkt. No. 567-21 (Meta employee stating, "I feel that using pirated material should be beyond our ethical threshold.").

[6] *Id.*

[7] U.S. Chamber of Commerce, Impacts of Digital Piracy on the U.S. Economy (June 2019), https://www.uschamber.com/technology/data-privacy/impacts-of-digital-piracy-on-the-u-s-economy.

[8] *Id.*

[9] The Authors Guild, Piracy, https://authorsguild.org/advocacy/piracy/, ("Each year, the publishing industry loses hundreds of millions of dollars in lost sales to piracy—and with each lost sale, authors lose royalty income.").

income.[10] Piracy does not just deprive rightsholders of the fruits of their labor, it also erodes morale and trust in the creative sector and normalizes theft of intellectual property, diminishing incentives for future innovation. The business models of entire creative industries are at risk.

When Generative AI Companies Ingest Pirated Materials, They Contribute to and Fuel the Proliferation of Shadow Libraries

Digital shadow libraries directly benefit from the ingestion activities of GenAI companies. When GenAI companies mine their work, they drive digital traffic to the libraries. Further, in at least one case, the shadow libraries derive direct benefits from GenAI companies. In one notable case, a shadow library known as Anna's Archives openly offers to work with AI companies in exchange for a "donation" or data trades. Contrary to the view of at least one district court judge, this offers to trade access to pirated materials for money and/or data indicates that pirate libraries can engage in symbiotic relationships with GenAI companies. In sum, the training of GenAI on materials can contribute to the proliferation and growth of digital piracy. These shadow libraries have continued to proliferate, with some of the largest such as Library genesis now claiming to have more than 2.4 million non-fiction books, 80 million science magazine articles and Anna's Archive with 36 million books and 103 million academic papers. [11]

Innovation Can Be Fostered, But Not at The Expense of Fair Compensation of Creative Economy Stakeholders and Support of Creative Markets

It is widely agreed that GenAI companies are the engines of innovation, and their emerging technologies hold great promise of enhancements in every area of life. None of the proposals raised here are intended to hamper such vital innovation. Indeed, none would hinder productive innovation: GenAI companies have the means to license uses of works just as every user of creative works has licensed uses since copyright was put into place. The music industry, to take just one example, is built on a system of rights clearances and licensing arrangements. Similarly, the film industry regularly engages in complex cross-licensing; as do the biotech, biomed, and pharma industries. Licensing works is standard business practice in every creative industry, and with good reason: it enables markets to operate efficiently and at optimal productivity.

Copyright itself exists to boost innovation, and to incentivize risk-taking in commercial markets. Innovation is the goal of copyright, as set forth in the Constitution: "to promote progress in Science and the Useful Arts." To the contrary, training LLMs on pirated works subverts the copyright system and fosters illicit activity that costs industries millions of dollars in revenues. This runs counter to innovation, as it disincentivizes creation, invention and discovery, and commercial productivity. GenAI companies do not need to train their models on illicit materials. There is already a thriving marketplace for the works they need to ingest; and they have the means to participate in the market for creative works just as every other user-consumer does on a daily basis.

---

[10] The Authors Guild, *Authors Guild Survey Shows Drastic 42 Percent Decline in Authors Earnings in Last Decade* (January 5, 2019), https://authorsguild.org/news/authors-guild-survey-shows-drastic-42-percent-decline-in-authors-earnings-in-last-decade/.

[11] https://greycoder.com/a-list-of-the-largest-shadow-libraries/

Generative AI Companies Are Arguably Engaged in Willful Acts That May Rise to the Level of Criminal Copyright Infringement

Training of GenAI on pirated materials promotes copyright infringement at two stages: the initial acts of digital piracy and the subsequent acts of training on materials under copyright without permission, licensing, or other licit forms of use. During the ingestion stage of these illicit practices, GenAI companies have knowingly, intentionally, and willfully chosen to circumvent copyright law and policy through their recourse to pirate repositories.

Historically, practices circumventing copyright protection have been successfully indicted on the basis of criminal copyright infringement, particularly where the actions were made with willful knowledge of such infringement.[12] Criminal copyright infringement requires a finding that copyright infringement was undertaken "willfully" and "for purposes of commercial advantage or private financial gain."[13] GenAI companies are engaging in ingestion of pirated materials with knowledge, constructive or actual, that the works on which they are building their LLM models are illicit sources. It is inarguable that they are acting for the purpose of commercial advantage. Therefore, as both elements are met, a strong argument can be made that their activities rise to the level of criminal copyright infringement.

Yet even if GenAI companies are not subject to criminal copyright infringement actions, the fact that they are clearly knowing, intentional, willful, and bad faith actors that resort to training on pirated materials should give a strong ground for denying them the ability to claim that their training is defensible under the fair use doctrine.

Congress Can Act by Ensuring That Generative AI Companies Adhere to Licensing Practices That Are Well-Established Practices in Copyright Law and Policy

By ingesting materials drawn from pirated repositories, GenAI companies are doing an end-run around copyright licensing, which is well-established under copyright law as the appropriate means of facilitating lawful access to, and use of, copyright owners' works.

For commercial markets in creative works to function fairly, sustainably, and optimally, licensing practices must be followed. GenAI companies must be required to follow these practices, as is required of all other participants in the creative markets. Courts have not yet offered a clear way to ensure that GenAI companies adhere to proper licensing practices; nor have they shown hold GenAI may be held accountable when they deviate from such well-settled practices. The time is ripe for Congressional action.

This is an area that urgently calls for guardrails and oversight. Congress can step in by requiring GenAI companies to limit their LLM training to legally-obtained and properly-licensed works. Some disclosure of training materials on the part of GenAI companies would allow oversight and, where necessary, course correction. These reasonable measures would simply bring GenAI

[12] United States v. Gordon, 37 F.4th 767 (1st Cir. 2022).
[13] https://www.justice.gov/archives/jm/criminal-resource-manual-1847-criminal-copyright-infringement-17-usc-506a-and-18-usc-2319

companies into line with standard and established licensing practices that exist in every commercial sector.

The time is ripe for Congress to make its voice heard. By requiring GenAI to follow fair and honest practices that are consistent with bedrock copyright laws and policies – including training its LLMs on materials acquired fairly and honestly, and engaging in well-established licensing practices – Congress can simultaneously foster innovative AI, support productive creators, and expand works available to the public. This can and should be a win-win for stakeholders in the technology industries, creative sectors, and the general public.