

AI Copyright Litigation Continues as NVIDIA Training Data Case Moves Forward

May 4, 2026

By W. Drew Kastner, Edward Baxter, Stephenie Wingyuen Yeung
楊穎苑 and LuAnne Morrow



A ruling earlier this month by Judge Jon S. Tigar in *Nazemian et al. v. NVIDIA Corp.*, No. 4:24 cv 01454 JST (N.D. Cal. filed Mar. 8, 2024), declining to dismiss key claims in the case following NVIDIA's motion to throw out portions of the complaint, signals that courts continue to be reluctant to resolve copyright disputes concerning AI training and outputs at the pleading stage. The ongoing class action against NVIDIA demonstrates why disputes over AI training data sourcing will continue to shape copyright doctrine well beyond the first wave of generative AI cases.

In *Nazemian* a class of authors, including Abdi Nazemian, Brian Keene, Stewart O'Nan, Susan Orlean, and Andre Dubus III, allege that Nvidia violated the Copyright Act by copying and storing unauthorized digital copies of their books to train its NeMo Megatron large language models, asserting claims for direct infringement, contributory and vicarious infringement, statutory damages, and injunctive relief. They also make claims under the Digital Millennium Copyright Act, alleging removal of copyright management information.

Central to the case are the plaintiffs' allegations that NVIDIA's training datasets incorporated pirated works sourced from "shadow libraries," including Books3 (derived from Bibliotik), The Pile, SlimPajama, and Anna's Archive, each of which allegedly contain massive numbers of unauthorized copies of copyrighted books. Unlike earlier AI disputes that focused on whether model outputs were substantially similar to copyrighted works, the *Nazemian* action frames infringement as complete at the point of copying of the inputs into the model when works were allegedly downloaded and retained, regardless of whether subsequent model training is transformative.

In allowing the direct infringement and related claims to proceed, the court made clear that fair use presents a mixed question of law and fact not suited for resolution on a Rule 12(b)(6) motion, particularly where the provenance, scope, and scale of the copied materials remain disputed. The ruling ensured that NVIDIA would not obtain an early exit from the litigation and underscored that allegations of unlawful data acquisition alone can carry a complaint past the pleading stage.

The *Nazemian* litigation sits within an expanding ecosystem of AI copyright cases, which at present comprises more than 50 such actions pending in U.S. federal courts, including actions involving Meta Platforms, Anthropic, and OpenAI. While recent fair-use rulings have not stemmed the AI litigation tide, the legal discussion has shifted from abstract debates about innovation policy to examinations of data sourcing, internal decision-making, and statutory compliance. Even as courts acknowledge that AI training may satisfy the "transformative use" inquiry, they continue to treat market harm, licensing markets, and unlawful acquisition as fact-dependent questions. It appears that so long as AI developers rely on massive training data sets and courts remain skeptical of practices involving pirated or unlicensed sources, copyright litigation over AI training models will continue to pervade.