Ballard Spahr

Novel Ruling Offers Framework for 'Fair Use' of Copyrighted Material for Training AI Systems

by Steven B. Winters and Harlan Mechling

July 2, 2025

Summary

The recent federal court finding—that using copyrighted books to train an AI large language model (LLM) qualifies as fair use—provides some guidance for companies developing or deploying generative AI systems and for businesses that rely on AI vendors.

The Upshot

- In *Bartz v. Anthropic*, the court found that LLM training is a "quintessentially transformative" use under the Copyright Act—marking a major development in the legal treatment of generative AI.

- However, the court also found that using or retaining "pirated," *i.e.*, unauthorized, copies of books for any purpose, including model training, prevents an alleged infringer from asserting the fair use defense.

- The court did not address whether AI-generated outputs infringe copyright.

- The decision from the Northern District of California is likely not the final word on these issues, as the ruling is subject to appeal and is just one of several relevant cases making their way through federal courts.

The Bottom Line

The ruling offers a practical framework for evaluating how courts may assess the sourcing, use, and storage of copyrighted materials in AI development. Companies building their own models should ensure that all training data is lawfully sourced. Businesses that license or integrate AI tools should ensure that contracts contain appropriate representations and risk allocation terms.

In *Bartz v. Anthropic*, Judge William Alsup issued a significant decision that delineates the copyright boundaries governing generative AI development. This is the first federal court order to meaningfully apply the fair use doctrine to AI training on copyrighted books, marking an important step in the evolving intersection of copyright law and artificial intelligence. The decision provides some guidance for companies developing or deploying generative AI systems—and for businesses that rely on AI vendors.

Bartz is a copyright infringement lawsuit brought by three authors against Anthropic PBC, the developer of the Claude generative AI system. The plaintiffs allege that Anthropic copied their books without authorization and used them in various ways, including building a central digital library and training LLMs.

The court evaluated Anthropic's conduct through a three-part framework, examining (1) the use of copyrighted works for model training, (2) format conversion of lawfully acquired books, and (3) the maintenance of a broad library of pirated content. Each use was evaluated separately under the four-factor fair use test.

Training on Copyrighted Books

The court determined that using copyrighted books to train Anthropic's large language models was "spectacularly" transformative and qualified as fair use. The training process was transformative because it did not involve copying the expressive content of

books for consumption or redistribution, but instead used the material to extract statistical patterns and relationships that enabled the model to generate new outputs. This function, the court explained, was fundamentally different from the original purpose of the works and aligned with how the fair use doctrine protects learning and innovation. The court also rejected the notion that a machine's use of copyrighted material should be treated less favorably than a human's, declining to draw a distinction based on the nature of the user.

In assessing market harm, the court found no evidence that Anthropic's use supplanted the market for the plaintiffs' books or posed a concrete risk of substitution. The plaintiffs did not allege that the Claude model produced infringing outputs, and the record contained no examples of the model generating content that reproduced or closely tracked their works. The court also noted that Anthropic had implemented filtering mechanisms aimed at preventing reproduction of copyrighted books. On this record, the court concluded that the training process did not interfere with the actual or potential market for the original works.

Digitizing Lawfully Purchased Books

The court found that Anthropic's conversion of purchased print books into digital files for internal use also qualified as fair use, though under a separate rationale. Anthropic had scanned physical books—destroying the originals in the process—to create searchable files stored in its internal research library. The court characterized this as a legitimate form of format-shifting, relying on precedent from *Sony Betamax*, *Texaco*, and *Google Books*. It emphasized that the digital copies were not distributed externally and functioned as one-to-one replacements for the original books. On these facts, the court concluded that the practice was "even more clearly transformative" than the uses upheld in those earlier cases.

Retention of Pirated Digital Books

The court reached a different conclusion with respect to Anthropic's handling of pirated digital books. It found that Anthropic had downloaded millions of unauthorized copies and retained them in its internal library, including materials that were no longer being used for model training. The court's holding turned on this indefinite retention for open-ended internal use, which it concluded was not a fair use under Section 107.

However, the court reasoned that even if the pirated books had been used solely to train an LLM and then immediately discarded, that use would still fall outside the fair use defense. In the court's view, fair use cannot excuse the initial unlawful acquisition of copyrighted material, regardless of whether the downstream application is transformative. This aspect of the ruling signals a rejection of the idea that pirated inputs can be justified by how they are later used in the AI development process.

Key Limitations

The court's analysis was limited to the input-side use of copyrighted books. Because the plaintiffs did not allege that Anthropic's models produced infringing outputs, the court did not address whether model-generated content could give rise to liability. It acknowledged that the Claude system included filtering mechanisms which prevented infringement of the authors' works that served as inputs. The court noted that Claude did not output to the public any further copies. This leaves open for future cases how copyright law might apply to LLM outputs.

* * *

The *Bartz* decision provides some guidance for companies developing or deploying generative AI systems—and for businesses that rely on AI vendors. Companies building their own models should ensure that all training data is lawfully sourced through licensing, purchase, or use of public domain content, and should implement internal controls to prevent the use or retention of unauthorized materials. Businesses that license or integrate AI tools from third parties should consider asking vendors about the provenance of training data and should ensure that contracts contain appropriate representations and risk allocation terms. The *Bartz* decision confirms that lawful sourcing is not merely best practice—it is a threshold requirement for defending infringement cases and asserting fair use in the context of AI training.

Ballard Spahr is closely monitoring developments at the intersection of copyright law and artificial intelligence. Our AI Legislation and Litigation Tracker provides a comprehensive view of AI-related legislative activities and important information about litigation matters with significant potential impact on clients.

This alert is a periodic publication of Ballard Spahr LLP and is intended to notify recipients of new developments in the law. It should not be construed as legal advice or legal opinion on any specific facts or circumstances. The contents are intended for general informational purposes only, and you are urged to consult your own attorney concerning your situation and specific legal questions you have.

Contacts

Steven B. Winters
OF COUNSEL

winterss@ballardspahr.com

206.223.7740

Harlan Mechling
ASSOCIATE

mechlingh@ballardspahr.com

503.778.2104