PUBLICATION

The Northern District of California Weighs in on "Transformative" Use of Copyrighted Books for Training Generative AI Tools

Authors: Edward D. Lanquist, Jr, Dominic A. Rota
June 27, 2025

Only a few months ago, the U.S. District Court for the District of Delaware ruled that the use of "headnotes" in a legal search tool, for the purpose of training a competing legal tool driven by artificial intelligence (AI), did not constitute "transformative" fair use. *See Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, 765 F. Supp. 3d 382 (D. Del. 2025).

While this order is on appeal to the Third Circuit, the U.S. District Court for the Northern District of California has weighed in on the issue of transformative use in the training of tools powered by generative AI, issuing two momentous decisions in a matter of days: *Andrea Bartz, et al. v. Anthropic PBC*, Civil Action No. 24-05417 WHA (N.D. Cal. Jun. 23, 2025) (*Anthropic PBC*) and *Richard Kadrey et al. v. Meta Platforms, Inc.*, Civil Action No. 23-cv-03417-VC (N.D. Cal. Jun. 25, 2025) (*Meta Platforms*).

The Northern District of California Has a Say on Fair Use

In *Anthropic PBC*, Judge William Alsup of the Northern District of California entered an order of summary judgment on the issue of fair use, finding in favor of the developer of the AI tool: Anthropic PBC (**Anthropic**). Only two days later, in *Meta Platforms*, Judge Vince Chhabria of the Northern District of California likewise issued an order in favor of the creator of another generative AI tool: Meta Platforms, Inc. (**Meta**).

Though *Anthropic PBC* and *Meta Platforms* present competing analyses on the application of the fair-use doctrine to generative AI, each court in the Northern District was tasked with the same question: whether copyright-protected textual works may be used to train an AI tool under the doctrine of fair use.

The Same Technology in Focus

Anthropic PBC and *Meta Platforms* examine copyright infringement through the lens of the same technology: generative AI. Generative AI consists of algorithms, or neural networks, that use training data to create new content in the form of text, images, or audio (See IBM's "What is Generative AI?"). These tools are driven by various mechanisms, such as deep learning models, large language models, natural language processing, and diffusion models, that scour over training data to generate new content.

Not only is generative AI the focus in *Anthropic PBC* and *Meta Platforms*, but also, the generative AI in each case is powered by a large language model (LLM). An LLM is a particular type of generative AI model that is designed to, **first**, understand language by analyzing statistical relationships among words in textual training data, and **second**, to create raw text by predicting what words are expected to fall in sequence. The most notable LLMs in the AI marketplace are Google's Gemini and OpenAI's ChatGPT.

A Brief Look into the Facts of Each Case

Anthropic is an AI software firm that was founded by former employees of OpenAI. Anthropic licenses its own offering of AI software, "Claude," which is trained on LLMs. In order to train the LLMs underlying Claude, Anthropic downloaded, for free, millions of digital copies of copyrighted works from pirate sites on the internet.

Anthropic did not just rely on pirated works; Anthropic purchased copyrighted books, going through the tedious process of tearing of the books' bindings, scanning every page, and storing them in digitized, text-searchable files. Each of these works was moved to a central library, which would theoretically be retained in perpetuity. From the central library, Anthropic could select various sets, or subsets, of the digitized, text-searchable works to train the LLM underlying Claude.

Meta, the owner and operator of several social-media platforms, namely, Instagram, Facebook, and Whatsapp, developed a series of LLMs under the name "Llama." Llama is offered for free to members of the public who intend to use it for non-commercial purposes; however, if intended for commercial purposes, only a few editions are available for free download. In order to train its LLM, Meta copied several textual works from "shadow libraries," which are online repositories that provide things like books, academic journal articles, music, or films *for free*, regardless of whether the foregoing media is protected by copyright. Each of these books were downloaded to the datasets for the purpose of training the Llama LLMs.

Not surprisingly, several authors filed suit against Anthropic and Meta in separate actions, asserting that Anthropic's and Meta's use of their copyright-protected books constituted infringement.

The Doctrine of Fair Use

The doctrine of "fair use" protects the public from lawsuits for permissible uses of copyrighted content, which would otherwise constitute infringement. Section 107 codifies the doctrine of "fair use" and establishes four factors for determining whether use of a copyrighted work is a fair use:

1. the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

Unsurprisingly, the factor most material in each of the decisions was the first factor, which asks whether the secondary use of the copyrighted material is "transformative." In other words, the first factor explores whether, and to what extent, the new work merely supersedes the original work, ***or***, if it adds something new with a further purpose or different character. *See Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 528 (2023).

In *Anthropic PBC*, the court examined the applicability of the "fair use" doctrine across two categories of uses of textual works: (1) uses for the purpose of training the LLM; and (2) uses for the purpose of creating a central library, *but not for* training the underlying LLM. In *Meta*, the court did not bifurcate its analysis like in *Anthropic*. Rather, the *Meta* court simply explored whether copying books for the purpose of training LLMs constitutes permissible fair use under Section 107.

Anthropic PBC: Fair Use for Some Uses but Not Others

In *Anthropic PBC*, the court held that the use of copyrighted books for the purpose of training LLMs constitutes fair use. Principally, the court found that the first factor, i.e., the purpose and character of the use, weighed in favor of Anthropic, as the use of word-level statistical relationships inside an LLM is a "spectacularly" transformative fair use. Though the nature of the copyrighted work weighed against a determination of fair use, the court was swift to note that the other factors, i.e., amount and substantiality and the effect of the use upon the potential market, weighed in favor of fair use, thus pushing the totality of the factors in the direction of fair use. Crucially, the accused AI system could not reproduce and deliver the copied work as an output to a prompting user.

While the court did hold that Anthropic's conversion of the lawfully purchased print copies into digital library copies constitutes permissible fair use, the court went the other direction for pirated copies – at least for those that were acquired and retained in the central library. Expressing doubt that any accused infringer could ever meet its burden of explaining why downloading pirated copies of works is reasonably necessary for permissible fair use, the court concluded that "[p]irating copies to build a research library without paying for it, and to retain copies should they prove useful for one thing or another, was its own use – and not a transformative one." The court found that all other factors weighed against fair use.

Lastly, the court analyzed whether copying of works for the purpose of creating a central library, but not for training, was permissible fair use. On this issue, the court did not grant summary judgment for Anthropic. As the record had not yet been amply developed to this issue, Anthropic was not entitled to "an order blessing all copying 'that Anthropic has ever made after obtaining the data.'" Thus, it is expected that this issue will ripen as more discovery is taken with respect to this issue, and it is likely to trigger a subsequent order from the court – assuming the parties do not settle as a result of the court's fair-use decision.

Meta Platforms: Fair Use Across the Board

Like in *Anthropic PBC*, the court in *Meta Platforms* held that using copyrighted books to train LLMs constitutes fair use. The court began its analysis by focusing on the first factor, stating that the copying of the books is a "transformative" use, at least in part because "an LLM's consumption of a book is different than a person's," as it "ingests text to learn 'statistical patterns' of how words are used together in different contexts." The court also noted that the generative AI tool can accomplish that which cannot be achieved by a hypothetical professor: it has the potential to exponentially multiply creative expression in a way that teaching individual people does not, including brainstorming a creative-writing project or writing source code.

As to the second factor, the court found such factor favors the plaintiffs, as the works at issue, i.e., books (e.g., novels, memoirs, and plays), are so highly expressive that they are worthy of copyright protection. In that same vein, the court was swift to note that the second factor "has rarely played a significant role in the determination of a fair use dispute," and so it turned to the third factor. Remarkably, even though the *entirety* of the copyrighted books was copied, the court stated that such *entire* copying was done in furtherance of the "transformative' purpose of the LLMs: to support enhanced training of the LLMs.

Unlike *Anthropic PBC*, the court in *Meta Platforms* explored the importance of the fourth factor, stating that it alone could be dispositive of the issue of the fair-use doctrine. In the context of generative AI, the court noted that there may be at least three ways in which copyright could harm the market for the protected works:

5. the LLM will regurgitate the copyright-protected works;

6. unauthorized copying of works for training purposes harms the market for licensing said works for that very purpose; and
7. the LLM will output works that compete with the original works, thereby potentially displacing the market of the original works

As to arguments one and two, the court noted that each argument fails: the LLMs cannot reproduce the works in their entirety, nor could a copyright-holder have cognizably realized licensing revenue from a use that is otherwise transformative. As to the third argument, the court left the door open for rightsholders. The court stated that the rapid generation of countless works could create competition with original works, the effect of which is market dilution by AI-generated competitors. But, even in discussing the potential perils of a marketplace dominated by AI-generated content, the court noted that the plaintiffs could not carry the day on the fourth factor, as they had hardly presented factual evidence that could raise a genuine dispute of fact. Consequently, the court was bound by the limits of the record and entered summary judgment in favor of Meta.

Limits on the Opinions

To set expectations, each of these decisions arise from motions for summary judgment in a single, though respected, district court. As such, opinions from other district courts in different jurisdictions are expected to issue in the months to come – some coming out differently, while others coming out the same.

Notably, each opinion only relates to the copying of books. Consequently, the decisions do not specifically address other copied works, such as content on websites and social-media platforms. Given that these sources may be digested without using pirated copies – to the extent that they are freely available – it is unclear whether such copying would be allowed. Likewise, as to music, it is unsettled whether a generative AI tool's license to certain musical works would permit it to exceed mere copying (e.g., creating wholly new musical works as derivatives of the copied musical works).

Practical Guidance

Though the decisions in *Anthropic PBC*and *Meta Platforms* are helpful roadmaps for guiding individuals and entities in the creation or deployment of AI tools, they are by no means the "law of the land." As it is expected that the Northern District of California's fair-use determinations will be appealed in the very same way the District of Delaware's *Ross* decision has been appealed, all that can presently be done is wait to see how federal appellate courts will address the issue of fair use in the context of training generative AI platforms. But, even if there is no clear law on the issue, individuals and entities can nevertheless implement certain practices to monitor infringement of copyrighted works. For example, copyright holders can presently audit digital piracy channels, take actions to preserve evidence of infringement (e.g., prompt queries and outputs), and evaluate protocols to license their copyright-protected content, so that they are in the driver's seat for any subsequent uses of their copyrighted works.

On the other end, individuals and entities looking to create, refine, or deploy AI tools should proceed with caution when using third-party works. Exemplary actions can be taken to guard against improper use of copyright-protected content, such as: documenting an inventory of all assets used for training an AI tool, confirming their provenance such that the source can be traced; implementing filters to block disallowed prompts and other outputs; and maintaining detailed logs of the foregoing practices, so as to be able to demonstrate compliance, should the practices be challenged as infringing.

The *AnthropicPBC* and *Meta Platforms* decisions push on the fundamental questions of copyright infringement in the AI era: is the ingestion of copyright-protected content an infringement, and if it is, is it a permissible fair

use? Or, is the output of copyright-protected content an infringement, and if it is, is it a fair use under the Copyright Act? Until sweeping legislation takes effect, or an order from this nation's highest court is issued, it will remain unpredictable as to how issues of copyright infringement, or fair use, should be analyzed in the context of an AI engine.

If you have questions about copyright enforcement, the impact of AI on intellectual property law, or you are in need of guidance on deploying or using an AI-backed tool, please reach out to Ed Lanquist or Dominic Rota, who are both members of Baker Donelson's Intellectual Property Team.

Ed Lanquist's legal practice focuses on patent, trademark, and copyright litigation, intellectual property counseling, trademark prosecution, and technology law. With more than 30 years of practicing intellectual property (IP) law, he has legal experience with a variety of IP issues. He can be reached at elanquist@bakerdonelson.com

Dominic Rota is a registered patent attorney who concentrates his practice in all areas of intellectual property, namely, patents, trademarks, and copyrights, and he counsels clients on a broad range of matters for emerging technologies, such as artificial intelligence (AI). He can be reached at drota@bakerdonelson.com.