AI and Copyright: How Lessons from Litigation Can Pave the Way to Licensing

DR. KIRTI GUPTA & **DR. ELIAS ILIN**
FEBRUARY 15, 2026, 12:15 PM 💬 1

"As the AI and copyright battles unfold in the courts and the fair use doctrine is put to the test, some economic and technical analysis can help shed the light on the fair use factors, providing the empirical evidence that has been lacking."

As the AI revolution accelerates and continues to reshape traditional business models, it has triggered a cascade of new legal, regulatory and policy challenges. At the forefront of these emerging issues are a growing number of high-stakes legal battles between content creators and major Generative AI (GenAI) companies behind large language models (LLMs). This article

examines key legal themes and critical questions arising from recent developments at the intersection of AI and Copyright law.

Dozens of lawsuits have been filed by authors, artists and music publishing companies challenging how LLMs use copyrighted material during the training process. Among the most notable active cases are _The New York Times v. OpenAI_, which centers on the use of journalistic content; _Disney v. Midjourney_, focused on visual works; and _Encyclopedia Britannica v. Perplexity AI_, involving literary material. In addition, class action suits brought by groups of book authors have targeted companies such as Anthropic, Meta, OpenAI, and Databricks. A comprehensive and regularly updated list of these cases is maintained by the Copyright Alliance. At the core of these disputes is a central legal question: Does training AI models on copyrighted content without explicit authorization constitute copyright infringement? This inquiry turns on whether such uses can be shielded by doctrines like "fair use" in the United States, or their equivalents in other jurisdictions. Courts are now being asked to navigate complex legal, economic, and technical dimensions of the fair use doctrine. Key issues under scrutiny include:

- the purpose, character, and extent of using copyrighted works for training,
- whether such use is transformative,
- the availability and impact of licensing regimes for training data, and
- whether AI-generated outputs compete with the original works in the marketplace.

These considerations ultimately raise broader questions about appropriate remedies – damages, injunctive relief, or policy interventions. As courts unpack various aspects of the fair use doctrine, expert analysis is increasingly required to interpret concepts such as market substitution, transformative use, and economic harm. This article synthesizes recent legal decisions and probes the unresolved technical and economic issues still facing courts. As these landmark cases progress, they are poised to shape the legal framework governing AI for years to come.

The Fair-Use Analysis: Recent Developments

The "fair use" doctrine is codified in Section 107 of the 1976 Copyright Act. It provides that "the fair use of a copyrighted work . . . is not an infringement of copyright" and lists four non-exclusive factors that must be considered in determining whether a particular use is fair:

(1) The purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit 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

The U.S. Copyright Office has also weighed in directly through its most recent Generative AI Training report, issued as part of its broader Copyright and AI initiative. The report concludes that compiling AI training datasets, such as through digitizing and aggregating copyrighted content, implicates the copyright holder's reproduction rights. It rejects a blanket application of the fair use defense for training AI models and is widely interpreted as favoring rights holders. Particularly significant is the report's emphasis on the fourth fair use factor: the effect on the potential market for or value of the original work. It finds that AI training involving copyrighted material may inflict substantial economic harm through lost sales, missed licensing opportunities, and dilution of market value due to a flood of AI-generated content, even when outputs merely imitate human-created styles. Adding further controversy, U.S. Copyright Office Director Shira Perlmutter was abruptly dismissed by the Trump administration just one day after the report's pre-publication – an unexpected move that has drawn scrutiny from legal and policy observers. The courts have increasingly been taking a different direction.

One of the first significant judicial rulings in *Thomson Reuters v. Ross Intelligence Inc.* set a critical precedent for the fair use defense in cases involving use of copyrighted material for AI training. When Thomson Reuters (Westlaw) sued its competitor, Ross Intelligence, for direct copyright infringement, the court's summary judgment ruled in favor of Thomson Reuters and rejected Ross's fair use defense. The court's fair use analysis, particularly its focus on the first and fourth factors, proved pivotal. For factor 1: purpose and character of the use, the court determined Ross's use was not transformative, and that it was a commercial endeavor aimed at creating a direct "market substitute" for Westlaw's offerings, lacking a distinct purpose or character. The court found factors 2 and 3 to be more favorable towards Ross, as the copied headlines weren't publicly revealed nor the nature of the copyrighted work deemed particularly original. Crucially, the factor 4: effect on the market, was deemed the "most important element" and strongly favored Thomson Reuters. The court recognized an obvious potential market for licensing copyrighted material for AI training, concluding that Ross's unauthorized use directly harmed this emerging market. Thus, this ruling focused on the market for input data used in AI training, creating a strong incentive for AI companies to secure licenses.

Two recent summary judgment decisions in the Northern District of California – *Bartz v. Anthropic PBC* (June 23. 2025) and *Kadrey v. Meta Platforms, Inc.* (June 25, 2025) further illuminated an evolving legal landscape, revealing both areas of consensus and significant divergence on the issue of whether unauthorized use of copyrighted works to train GenAI LLMs is infringement or fair use.

In _Bartz v. Anthropic PBC_, a group of authors initiated legal action against Anthropic, alleging copyright infringement. The core of their claim was that Anthropic had unlawfully copied their books and used these copies to train its Claude AI model, incorporating both scanned copies of legitimately purchased books and millions of others allegedly downloaded from pirated databases such as PiLiMi and LibGen. Judge William Alsup's recent decision delivered a nuanced, split verdict. Applying fair use factors 1-3, a significant aspect of his ruling was the determination that Anthropic's copying of works specifically for the purpose of _training_ its LLMs constituted fair use, as he found such use to be "exceedingly transformative" and concluded that the fair use factors weighed in Anthropic's favor. Importantly, Judge Alsup ruled that it was _not_ fair use for Anthropic to use pirated copies for use in its central library. Shortly afterwards, Anthropic followed-up with one of the first large settlement proposals in such cases, offering $1.5 billion to the book authors which has recently been approved by Judge Alsup on a preliminary basis after raising some questions about the number of class members and who is included in the class.

Judge Alsup's decision applied factor 4 for determining effect on the market from the output of LLM models. This also weighed in Anthropic's favor, rejecting the plaintiffs' argument that training LLMs would lead to an explosion of competing works as lacking in any empirical evidence. Overall, the _Bartz_ decision represents a significant victory for AI companies seeking to establish that copying works to train LLMs can be fair use, especially when the data is lawfully acquired. It also emphasizes the importance of market evidence in claiming that LLM model outputs would compete with creators' works.

In a contemporaneous lawsuit, _Kadrey v. Meta Platforms, Inc._, a group of authors filed a complaint against Meta, alleging unauthorized copying of their books for the purpose of training Meta's LLaMA models. Similar to the allegations in _Bartz_, the plaintiffs claimed that their books were included in pirated, shadow libraries that Meta utilized for its AI training. Meta asserted fair use as its defense against these copyright infringement claims. On June 25, 2025, Judge Vince Chhabria issued a summary judgment decision that largely favored Meta. Judge Chhabria found Meta's use of the plaintiffs' works for training to be highly transformative and, based on the record presented, concluded that it qualified as fair use.

Judge Chhabria's decision considered factor 4 as a pivotal. Taking into account the effects on both the market for input data and on the market for the outputs, he weighted both in Meta's favor. 1) Market for input data: Judge Chhabria rejected the plaintiffs' argument that the unauthorized use harmed a potential licensing market for AI training data, deeming it a circular argument – he concluded that for a licensing market to exist, the court would have to assume that the use of training data is _not transformative_, without applying the fair use analysis to determine whether or not this is the case.. 2) Harm to market from output data: Judge Chhabria was highly receptive to the argument of indirect substitution and market dilution – the idea that a surge of AI outputs could saturate the market and compete with original works. He acknowledged LLMs'

unique ability to rapidly create millions of secondary works that could compete with original works. He even stated that "it seems likely that market dilution will often cause plaintiffs to decisively win the fourth factor — and thus win the fair use question overall." However, he ultimately ruled for Meta on this factor because the plaintiffs failed to present any evidence of this type of harm.

Economic and Technical Analysis: Importance of Evidence

Emphasizing the importance of the fact-specific nature of fair use and of empirical evidence, these recent decisions provide critical nuances for the types of potential technical, economic, and statistical analyses that are required in AI copyright litigation. We list these below.

(1) Inputs: User Prompt Analysis

Sophisticated statistical analysis can be applied to determine how transformative the GenAI models are in taking the creative inputs as training data and producing outputs. Statistical and text-matching models of large sets of user prompts can determine how users interface with models, what use cases are most common, and to what extent there is any overlap between inputs used as training data and outputs generated from the model. Such analysis is typically done by statistical experts who have interpreted large data sets, identified broad patterns and outliers, with a robust knowledge of GenAI models and how they function.

A notable example of the need for such analysis is the recent *New York Times v. OpenAI* case, where the New York Times (NYT) alleged that Open AI's models were generating outputs that were too similar to the original training data, raising concerns about copyright infringement and the transformative nature of the AI-generated content. In this case, a thorough statistical analysis of the inputs and outputs of the Open AI models could help determine the extent to which the models were truly generating novel content or simply reproducing existing work. By applying statistical and text-matching models, experts can identify patterns and anomalies in the data, such as instances of over-reliance on specific training data sources, or cases where the model's outputs were excessively similar to the original inputs.

(2) Inputs: Cost of Infringement

Courts have also considered whether it is practically feasible for GenAI models to remove any specific data from their corpus of training data, and at what cost (typically in the context of injunctive relief). GenAI models are now trained on terabytes of raw data, translating into billions of tokens. Any training dataset is translated into tokens which are fed into creating models' parameters and weights. Once the model is already "trained", it may not be

possible to "untrain" the model without using the data. In order to determine the feasibility and cost of options, computer scientists with expertise in machine learning GenAI training models can do a fact specific analysis.

Indeed, in the *Bartz v. Anthropic PBC* matter, Judge Alsup alluded to the cost of removing the training data. Both parties agreed to certain guardrails Anthropic put in place to ensure that the output from the models will not regurgitate any content from the inputs.

(3) Inputs: Licensing Landscape

The availability and impact of licensing regimes for training data is another complex aspect of the fair use doctrine that courts must face. GenAI models are trained on an increasing amount of data with multiple inputs. For example, Chat-GPT's GPT-4 data set is trained <u>on one petabyte of data</u>. For this reason, any individual or collected work represents a fraction of the training data, and it is difficult to determine the incremental value of the input data for the purposes of training. In addition, a potential market for licensing of such data is fraught with several difficulties. First, with rights for creative works fragmented and distributed across multiple rights holders leads to significant transaction costs for aggregating the volume of required permissions and negotiating licensing rights. Second, there is incomplete information about who the rights owners are and the value of such data. On the other hand, there have been <u>collective licensing agreements</u> in the past for fragmented copyrighted materials such as music. More recently, several licensing agreements for training data used by GenAI models <u>have emerged</u> that may serve as comparable licenses in some circumstances, *if* the data and use cases are similar.

(4) Outputs: Evidence of Harm to Market

The U.S, Copyright Office's report as well as the recent *Kadrey v. Meta Platforms, Inc.* and *Bartz v. Anthropic PBC* cases have emphasized the fourth factor of the fair use analysis. Fair use factor 4 requires evidence of harm to the market for the creative works due to a potential competition or substitution effect that may be created by any output from the GenAI models that are using the creative works as input training data. Such evidence, however, is hard to produce, especially, if the analysis of the output produced by most user prompts from the GenAI models is transformative enough and is not being used to recreate content. In addition, further market survey analysis can be used for specific creative works to determine what attributes drive the demand for creative content. For example, consumers may be buying a music album or a song largely due to the brand name of a singer, and not the songs' lyrics.

So far, the courts have not been persuaded by the evidence provided by the plaintiffs. In *Kadrey v. Meta Platforms, Inc.*, Judge Chhabria acknowledged LLMs' unique ability to rapidly create millions of secondary works that could compete with original works but ultimately ruled in Meta's favor as plaintiffs failed to present any evidence of this type of harm. In *Bartz v. Anthropic PBC*, Judge Alsup's decision also weighed in Anthropic's favor, rejecting the plaintiffs' argument that training LLMs would lead to an explosion of competing works due to lack of any empirical evidence.

Economic Analyses Can Help Forge the Path Ahead

As the AI and copyright battles unfold in the courts and the fair use doctrine is put to the test, some economic and technical analysis can help shed the light on the fair use factors, providing the empirical evidence that has been lacking. For example, user prompt and input-output analysis can help to determine how transformative the output of a GenAI model is. The effect on the market for licensing of the input data as well as the effect on the market for copyrighted content from the outputs generated from the models can be analyzed with economic analysis, growing empirical evidence, and market surveys. The cost of potentially removing any specific data points from a corpus of training data is also a technical and economic question. All of these can help in the liability assessment and damages analyses of the ongoing cases.

DR. KIRTI GUPTA

Kirti Gupta is Vice President and Chief Economist of Global Technology at Cornerstone Research. She is an engineer and economist specializing in technology; intellectual property (IP) issues such as standard [...see more]

DR. ELIAS ILIN

Dr. Elias Ilin is Associate at Cornerstone Research and specializes in intellectual property and antitrust matters. He holds a Ph.D. in Economics from Boston University and a dual Bachelor's degree [...see more]

A **pre-**publication item has exactly zero level of persuasion as to having a policy directive.

Not only is the legal question of *how* policy is put forth *any* position tied to the ongoing legal battle (policy is most often an Executive Branch function), the status of 'pre-publication' bars anyone from reliance on the views expressed therein.

Those views do not have any persuasive authority.

At IPWatchdog.com our focus is on the business, policy and substance of patents and other forms of intellectual property. Today IPWatchdog is recognized as the leading sources for news and information in the patent and innovation industries.

Terms & Conditions of Use │ Privacy Policy

Images on IPWatchdog Primarily
Provided by