The Indian EXPRESS
JOURNALISM OF COURAGE

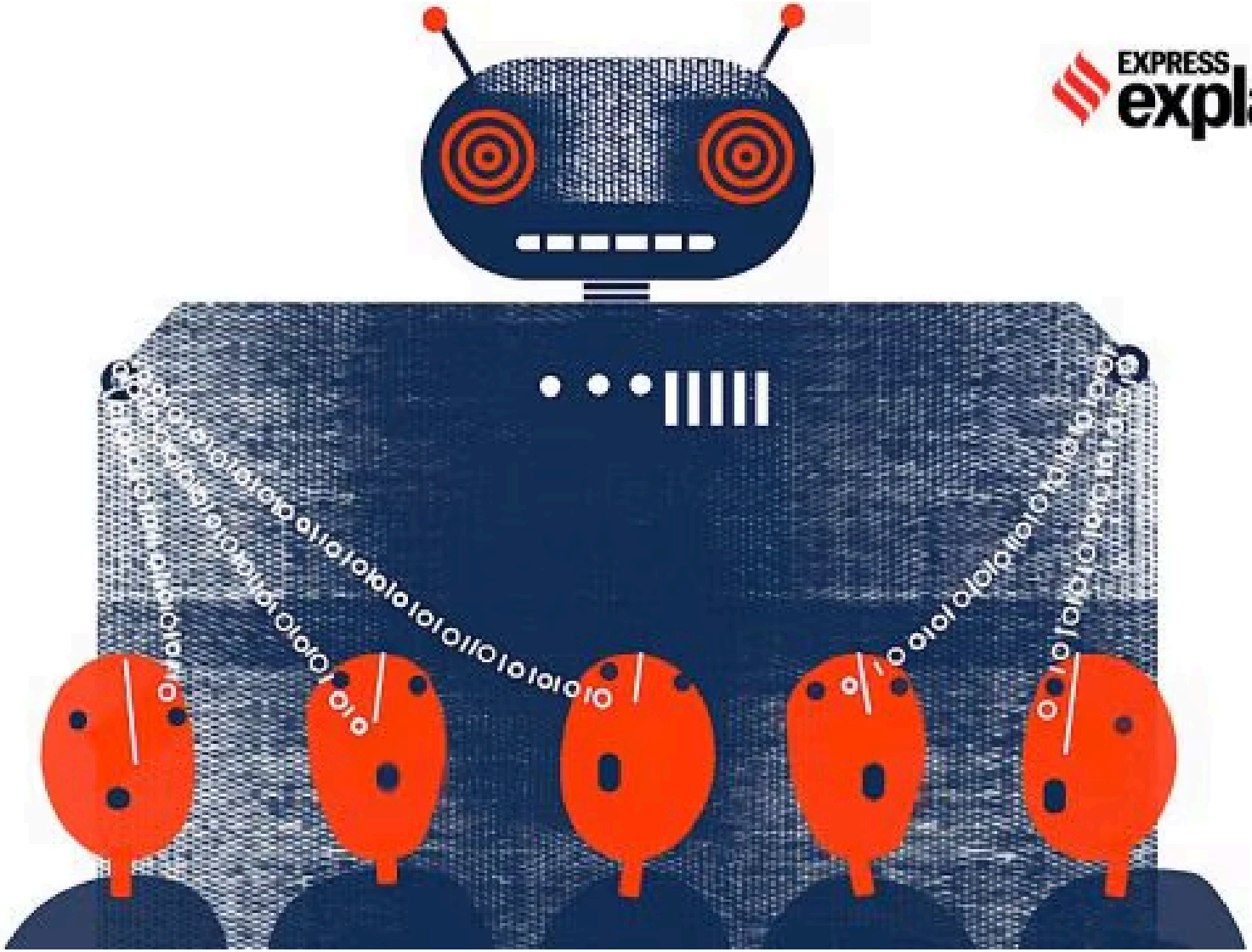
# Explained: AI & copyright law

Are generative artificial intelligence models built on stolen creative work? The first two judgments addressing this question in US courts have sided with tech companies. But the matter is far from settled

Written by Vidhatri Rao
New Delhi | July 3, 2025 07:12 IST



6 min read



There are at the moment at least 21 ongoing lawsuits in the US, filed by writers, music labels, and news agencies, among others, against tech companies for training AI models on copyrighted work. This, the petitioners have argued, amounts to "theft".

In two key copyright cases last week, US courts ruled in favour of tech companies developing artificial intelligence (AI) models.

While the two judgments arrived at their conclusions differently, they are the first to address a central question around generative AI models: are these built on stolen creative work?

At a very basic level, AI models such as ChatGPT and Gemini identify patterns from massive amounts of data. Their ability to generate passages, scenes, videos, and songs in response to prompts depends on the quality of the data they have been trained on. This training data has thus far come from a wide range of sources, from books and articles to images and sounds, and other material available on the Internet.

There are at the moment at least 21 ongoing lawsuits in the US, filed by writers, music labels, and news agencies, among others, against tech companies for training AI models on copyrighted work. This, the petitioners have argued, amounts to "theft".

In their defence, tech companies say they are using the data to create "transformative" AI models, which falls within the ambit of "fair use" — a concept in law that permits use of copyrighted material in limited capacities for larger public interests (for instance, quoting a paragraph from a book for a review).

Here's what happened in the two cases, and why the judgments matter.

**Case 1: Writers vs Anthropic**

In August 2024, journalist-writers Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson filed a class action complaint — a case that represents a large group that could be/were similarly harmed — against Anthropic, the company behind the Claude family of Large Language Models (LLMs).

The petitioners argued that Anthropic downloaded pirated versions of their works, made copies of them, and "fed these pirated copies into its models". They said that Anthropic has "not compensated the authors", and "compromised their ability to make a living as the LLMs allow anyone to generate — automatically and freely (or very cheaply) — texts that writers would otherwise be paid to create and sell".

Anthropic downloaded and used Books3 — an online shadow library of pirated books with about seven million copies — to train its models. That said, it also spent millions of dollars to purchase millions of printed books and scanned them digitally to create a general "research library" or "generalised data area".

Judge William Alsup of the District Court in the Northern District of California ruled on June 23 that Anthropic's use of copyrighted data was "fair use", centering his arguments around the "transformative" potential of AI.

Alsup wrote: "Like any reader aspiring to be a writer, Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them — but to turn a hard corner and create something different. If this training process reasonably required making copies within the LLM or otherwise, those copies were engaged in a transformative use."

**Case 2: Writers vs Meta**

Thirteen published authors, including comedian Sarah Silverman and Ta-Nehisi Coates of Black Panther fame, filed a class action suit against Meta, arguing they were "entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law".

The thrust of their reasoning was similar to what the petitioners in the Anthropic case had argued: Meta's Llama LLMs "copied" massive amounts of text, with its responses only being derived from the training dataset comprising the authors' work.

Meta too trained its models on data from Books3, as well as on two other shadow libraries — Anna's Archive and Libgen. However, Meta argued in court that it had "post-trained" its models to prevent them from "memorising" and "outputting certain text from their training data, including copyrighted material". Calling these efforts "mitigations", Meta said it "could get no model to generate more than 50 words and punctuation marks…" from the books of the authors that had sued it.

In a ruling given on June 25, Judge Vince Chhabria of the Northern District of California noted that the plaintiffs were unable to prove that Llama's works diluted their markets. Explaining market dilution in this context, he cited the example of biographies.

If an LLM were to use copyrighted biographies to train itself, it could, in theory, generate an endless number of biographies which would severely harm the market for biographies. But this does not seem to be the case thus far.

However, while Chabbria agreed with Alsup that AI is groundbreaking technology, he also said that tech companies who have minted billions of dollars because of the AI boom should figure out a way to compensate copyright holders.

Significance of rulings

These judgments are a win for Anthropic and Meta. That said, both companies are not entirely scot-free: they still face questions regarding the legality of downloading content from pirated databases.

Anthropic also faces another suit from music publishers who say Claude was trained on their copyrighted lyrics. And there are many more such cases in the pipeline.

Twelve separate copyright lawsuits filed by authors, newspapers, and other publishers — including one high-profile lawsuit filed by The New York Times — against OpenAI and Microsoft are now clubbed into a single case. OpenAI is also being separately sued by publishing giant Ziff Davis.

A group of visual artists are suing image generating tools Stability AI, Runway AI, Deviant Art, and Midjourney for training their tools on their work. Stability AI is also being sued by Getty Images for violating its copyright by taking more than 12 million of its photographs.

In 2024, news agency ANI filed a case against OpenAI for unlawfully using Indian copyrighted material to train its AI models. The Digital News Publishers Association (DNPA), along with some of its members, which include The Indian Express, Hindustan Times, and NDTV, later joined the proceedings. Going forward, this is likely to be a major issue in India too.

Thus, while significant, the judgments last week do not settle questions surrounding AI and copyright — far from it.

And as AI models keep getting better, and spit out more and more content, there is also the larger question at hand: where does AI leave creators, their livelihoods, and more importantly, creativity itself?



This article went live on July third, twenty twenty-five, at twelve minutes past seven in the morning.

---

**TAGS:** **Artificial Intelligence** **Copyright** **Copyright Act**