New York Times Fights OpenAI's 'Unprecedented' Bid for Journalistic Materials

aibusiness.com/nlp/new-york-times-fights-openai-s-unprecedented-bid-for-journalistic-materials

Ben Wodecki, Jr. Editor
July 10, 2024

4 Min Read

Gary Hershorn/Getty Images

The copyright battle between The New York Times and OpenAI intensified as the AI company pushed for access to journalists' notes and memos in a contentious discovery dispute.

The Times sued OpenAI in December 2023, contending that ChatGPT was trained on its articles without permission, with the chatbot able to generate “near-verbatim excerpts” from its articles.

As the dispute develops, OpenAI is trying to access materials it argues are crucial to its defense as part of the pretrial phase of the lawsuit.

The Times, however, is trying to prevent OpenAI from accessing what it considers to be an “overbroad and unduly burdensome” amount of material “not relevant to any party’s claims or defenses.”

OpenAI is demanding copies of the related reporter’s notes, interview memos and records of materials and articles claimed to have been copied by ChatGPT.

The AI developer argued in a July 1 memo to U.S. District Judge Sidney H. Stein that its discovery demands were relevant to both the Times' infringement claims and OpenAI's defenses, such as fair use.

OpenAI argued that simply providing the actual works and not the wider related materials that went into creating them would be "insufficient" to allow them to test the assertions that the articles are the Times's "original works of authorship" entitled to copyright protection.

“The Times refuses to produce the vast majority of documents sought in these requests, agreeing only to produce the actual works at issue,” the memo reads. “OpenAI cannot determine from the works alone which portions reflect human-authored content original to the Times and which portions do not.”

The Times responded in a memo published on July 3, arguing that OpenAI's request for reporters' notes and memos is “unprecedented and turns copyright law on its head.”

The newspaper’s lawyers contend that the discovery demands were irrelevant and served “no purpose other than harassment and retaliation.”

“The Times news gathering process is not on trial,” the memo reads. “OpenAI and Microsoft's infringement of millions of the Times's registered copyright works is."

The Times contends that even if an article contains mostly quotes, it would still be protected by copyright law, arguing the “expressive nature of a work is determined by reference to the work itself.”

Regarding allegations of infringement, the Times argues that only actual judicial determinations of infringement are relevant, not unsubstantiated claims.

“OpenAI has the burden to show that the discovery it seeks 'is more than merely a fishing expedition.' It has not done so,” the memo reads.

The legal battle between the two has become a high-profile clash concerning the fair use of copyright materials in AI training.

The heart of the issue is that OpenAI allegedly scraped articles from the New York Times without permission to train the underlying AI models for ChatGPT.

Data scraping in AI was once commonplace, with developers desperate to gather as much information as possible for their AI systems, arguing it was fair use, which applies protections for the use of copyright materials in the creation of new “transformative” works.

The practice of data scraping has become increasingly criticized lately, with rightsholders becoming protective of their content. Platforms are instead forcing developers to strike lucrative licensing deals to access their data. Only last month, Reddit blocked AI web crawlers in a bid to protect its data.

Several smaller regional news outlets followed the New York Times' lead, filing suit against OpenAI for "theft" of its copyrighted content.

OpenAI has routinely disputed the Times' infringement allegations, contending that the paper 'hacked' ChatGPT and manipulated it into generating copyrighted content.

The AI developer previously claimed it would not be able to develop state-of-the-art models without access to copyrighted material. The Microsoft-backed company has since struck licensing deals with News Corp, Stack Overflow and Axel Springer to secure access to training data.
