OpenAI desperate to avoid explaining why it deleted pirated book datasets

OpenAI may soon be forced to explain why it deleted a pair of controversial datasets composed of pirated books, and the stakes could not be higher.

At the heart of a class-action lawsuit from authors alleging that ChatGPT was illegally trained on their works, OpenAI's decision to delete the datasets could end up being a deciding factor that gives the authors the win.

It's undisputed that OpenAI deleted the datasets, known as "Books 1" and "Books 2," prior to ChatGPT's release in 2022. Created by former OpenAI employees in 2021, the datasets were built by scraping the open web and seizing the bulk of its data from a shadow library called Library Genesis (LibGen).

As OpenAI tells it, the datasets fell out of use within that same year, prompting an internal decision to delete them.

But the authors suspect there's more to the story than that. They noted that OpenAI appeared to flip-flop by retracting its claim that the datasets' "non-use" was a reason for deletion, then later claiming that all reasons for deletion, including "non-use," should be shielded under attorney-client privilege.

To the authors, it seemed like OpenAI was quickly backtracking after the court granted the authors' discovery requests to review OpenAI's internal messages on the firm's "non-use."

In fact, OpenAI's reversal only made authors more eager to see how OpenAI discussed "non-use," and now they may get to find

out all the reasons why OpenAI deleted the datasets.

Last week, US district judge Ona Wang ordered OpenAI to share all communications with in-house lawyers about deleting the datasets, as well as “all internal references to LibGen that OpenAI has redacted or withheld on the basis of attorney-client privilege.”

According to Wang, OpenAI slipped up by arguing that “non-use” was not a “reason” for deleting the datasets, while simultaneously claiming that it should also be deemed a “reason” considered privileged.

https://arstechnica.com/tech-policy/2025/12/openai-desperate-to-avoid-explaining-why-it-deleted-pirated-book-datasets/