# Author exposes widespread AI training copyright breaches and warns of cultural loss



In 2023, nine books authored by an unnamed writer, along with approximately 180,000 others, were reportedly subjected to copyright infringement after being unlawfully scraped from a pirate website for inclusion in the Books3 dataset, which is used to train large language models (LLMs) by major tech companies. This issue has been brought to light following a significant breach involving Meta, which allegedly accessed 7 million books from the LibGen pirate site, where the author's work, including foreign editions, was also found.

According to the author, they have come to the unsettling conclusion that the copyright status of their life’s work has been compromised within just a few years, primarily due to actions taken by tech companies conducting AI training. They highlighted findings from Alex Reisner’s investigation published in The Atlantic, suggesting that employees at these tech firms were aware of the illegality of their actions, with "deception preceding theft and... following theft."

Despite the distress caused by these violations, the author has sought to explore potential solutions. They have engaged with a licensing company, establishing a process to designate that no books authored by them may be exploited for AI purposes. This involves inputting details for 216 titles and setting up exclusions across two categories of AI usage. However, the author described this task as daunting, stating, "it’s like spitting into a strong wind," given the previous collusion between companies and pirate sites that facilitated such breaches.

The author reiterated a significant concern: once content has been scraped to train AI, it cannot be "unlearned." This complicates any efforts to reclaim rights, as the base material has already been integrated into AI models. They expressed frustration with the British government’s recent proposal that the training of AI on copyrighted works should largely fall under "fair usage." The proposal suggests that to stop their work from being used for AI training, authors must "opt out" every edition in every territory, which could prove unmanageable and impractical for many.

Moreover, the recent legislation appears to enable businesses to copyright derivatives produced from the training data, a point the author finds particularly troubling. They argue that this effectively endorses copyright infringement of existing works while simultaneously protecting the derivatives born from that infringement.

In lieu of concentrating on new fiction, the author has shifted their focus towards advocating against this proposed legislation, emphasising its inequity and potential repercussions for the literary community and cultural heritage. They contend that not only is there a financial aspect at stake, but crucial elements of humanity's capacity for abstract thought and storytelling are also under threat.

In a positive development, the House of Lords in the UK has reportedly called for amendments to the legislation, seeking improvements that would require AI crawlers to acknowledge UK copyright law and inform creators if their work has been scraped. Yet, the author lamented that discussions seem predominantly centred on economic concerns rather than the intrinsic value of creative works.

Ultimately, the author expressed the belief that the stakes are higher than mere revenue, arguing for the preservation of cultural narratives that embody wisdom across generations, while also raising concerns about the broader societal implications of technology on truth and knowledge.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.theguardian.com/australia-news/2023/sep/28/australian-books-training-ai-books3-stolen-pirated> - This article discusses the alleged use of thousands of Australian books, including works by Richard Flanagan, in the Books3 dataset for AI training, highlighting the widespread nature of the issue.
2. <https://www.theguardian.com/technology/2025/jan/10/mark-zuckerberg-meta-books-ai-models-sarah-silverman> - This piece reveals that Meta's CEO, Mark Zuckerberg, approved the use of pirated books from LibGen to train AI models, indicating internal awareness of the actions taken.
3. <https://www.theguardian.com/technology/2025/jan/10/mark-zuckerberg-meta-books-ai-models-sarah-silverman> - The article also highlights that Meta employees were aware of the pirated status of the books used for AI training, supporting the claim of internal knowledge of the illegality of their actions.
4. <https://www.theguardian.com/technology/2025/jan/10/mark-zuckerberg-meta-books-ai-models-sarah-silverman> - The piece discusses the ethical dilemma faced by Meta in using pirated resources for AI training, aligning with the author's concerns about the compromise of their life's work.
5. <https://www.theguardian.com/technology/2025/jan/10/mark-zuckerberg-meta-books-ai-models-sarah-silverman> - The article touches upon the broader implications of AI training on copyrighted works, resonating with the author's concerns about the societal impact of such practices.
6. <https://www.theguardian.com/technology/2025/jan/10/mark-zuckerberg-meta-books-ai-models-sarah-silverman> - The piece also discusses the challenges authors face in protecting their works from unauthorized AI training, echoing the author's frustration with the British government's proposal on 'fair usage'.
7. <https://www.deseret.com/opinion/2025/05/03/ai-engines-used-my-books-without-permission/> - Please view link - unable to able to access data