# Meta’s use of pirated books to train AI sparks global copyright backlash



The unfolding drama surrounding Meta and its alleged use of pirated books to train artificial intelligence systems has sparked significant debate among authors, technologists, and legal experts alike. Recent revelations indicate that Meta's AI language model, Llama, was reportedly trained on materials sourced from Library Genesis (LibGen), a notorious online repository of pirated content. This not only raises ethical questions but also ignites frustration within the literary community, already grappling with the uncertainties of digital copyright.

Back in January 2025, prominent authors including Ta-Nehisi Coates and Sarah Silverman filed a lawsuit against Meta, asserting that the tech giant knowingly utilised pirated works, as evidenced by internal documents that suggested awareness of the illegitimate status of these materials. These revelations followed a report by The Atlantic in March detailing how Meta opted for the inexpensive and efficient yet ethically questionable option of using pirated material rather than engaging in a slow and costly legal acquisition process. This decision highlights an ongoing struggle in the tech industry regarding the boundaries of intellectual property and the extent to which companies can exploit creative content for training AI models.

Notably, the fraudulent use of these materials has led to a domino effect of legal repercussions. In addition to the high-profile lawsuit from U.S. authors, French publishers took a stand, initiating their own legal actions against Meta. The implications extend beyond the litigation arena, as the situation raises significant concerns about the future of content creation and authors’ rights within a rapidly evolving digital landscape.

While the incident initially provoked outrage, it has led to a split perspective among authors regarding the implications of AI training on their work. For some, like the author of a recent commentary who found their reference book included in the dataset, the situation elicits a mix of emotions. They expressed curiosity and even validation at having their work acknowledged, indicating a willingness to rationalise the use of reference texts in this context. This author suggested that, as AI technology advances, it may be preferable for AI to draw from accurate, researched content rather than fabricating misinformation.

However, this sentiment shifts dramatically when it comes to genuinely creative works. The author is adamant that there is a distinct line between the use of data-rich reference texts and the appropriation of deeply creative expressions such as literature, music, or visual art. The concern is that AI's ability to emulate stylistic nuances presents a risk akin to identity theft. The uniqueness of creative works lies in their artistic expression, and mimicking that essence without proper attribution or compensation undermines the entire value system of the creative industries.

The legal landscape surrounding these issues is murky, with a strong likelihood that the precedent set by these lawsuits will influence how AI companies approach content sourcing in the future. A significant conflict remains as tech companies often invoke “fair use” and “transformative works” to justify their actions, despite resistance from creators who argue this rationale is misapplied in the context of AI-generated content. The potential outcome could lead to broader implications for copyright law, particularly as the landscape shifts towards AI as a predominant creator of content.

As the situation continues to evolve, questions remain about the future of both AI development and the rights of creators. A push for ethical licensing and fair compensation is gaining traction, with organisations like the Independent Book Publishers Association voicing their support for fair treatment of authors. Ultimately, navigating this complex terrain will require a balancing act—one that respects the rights of creators while addressing the realities of a technology-driven era.

The irony looms that, as AI-generated content fills our digital spaces, the cycle of content may eventually lead back to itself, resulting in a regurgitation devoid of true creativity. The landscape of artistic expression risks becoming an echo chamber, where originality is diluted in a relentless pursuit of algorithm-driven output. As the discussions unfold, it is clear that the path forward will demand scrutiny regarding ethical practices and a commitment to preserving the sanctity of creative expression in an increasingly automated world.

## Reference Map:

* Paragraph 1 – [[1]](https://www.creativebloq.com/ai/call-me-a-traitor-but-i-dont-mind-ai-being-trained-on-my-content-even-after-meta-stole-my-book), [[2]](https://www.reuters.com/technology/artificial-intelligence/meta-knew-it-used-pirated-books-train-ai-authors-say-2025-01-09/), [[3]](https://www.ft.com/content/b1f4965f-6ea6-4afd-968b-76bb2f7acfc4)
* Paragraph 2 – [[4]](https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/?utm_source=apple_news), [[5]](https://apnews.com/article/168b32059e70d0509b0a6ac407f37e8a), [[6]](https://www.ibpa-online.org/news/697191/IBPA-Statement-on-Metas-Use-of-Pirated-Books-to-Train-AI.htm)
* Paragraph 3 – [[1]](https://www.creativebloq.com/ai/call-me-a-traitor-but-i-dont-mind-ai-being-trained-on-my-content-even-after-meta-stole-my-book), [[7]](https://arstechnica.com/tech-policy/2025/02/meta-torrented-over-81-7tb-of-pirated-books-to-train-ai-authors-say/)
* Paragraph 4 – [[3]](https://www.ft.com/content/b1f4965f-6ea6-4afd-968b-76bb2f7acfc4), [[6]](https://www.ibpa-online.org/news/697191/IBPA-Statement-on-Metas-Use-of-Pirated-Books-to-Train-AI.htm)
* Paragraph 5 – [[2]](https://www.reuters.com/technology/artificial-intelligence/meta-knew-it-used-pirated-books-train-ai-authors-say-2025-01-09/), [[5]](https://apnews.com/article/168b32059e70d0509b0a6ac407f37e8a)

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

## Bibliography

1. <https://www.creativebloq.com/ai/call-me-a-traitor-but-i-dont-mind-ai-being-trained-on-my-content-even-after-meta-stole-my-book> - Please view link - unable to able to access data
2. <https://www.reuters.com/technology/artificial-intelligence/meta-knew-it-used-pirated-books-train-ai-authors-say-2025-01-09/> - In January 2025, Reuters reported that authors, including Ta-Nehisi Coates and Sarah Silverman, filed a lawsuit against Meta Platforms, alleging the company used pirated copyrighted books to train its AI systems. Internal Meta documents indicated awareness of the pirated status of the works, with the plaintiffs asserting that Meta utilized the AI training dataset known as LibGen, containing millions of pirated works, to train its large language model, Llama. The case highlights the ongoing debate over AI companies' use of copyrighted material for model training.
3. <https://www.ft.com/content/b1f4965f-6ea6-4afd-968b-76bb2f7acfc4> - A Financial Times article from April 2025 discusses a significant legal challenge Meta faces, defending itself against a lawsuit filed by U.S. authors, including Ta-Nehisi Coates and Richard Kadrey. The authors allege that Meta unlawfully used copyrighted books from the shadow library LibGen to train its Llama AI models. The case is one of the first major tests to determine whether AI companies can claim 'fair use' of copyrighted material for model training, especially when sourced from pirated databases.
4. <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/?utm_source=apple_news> - In March 2025, The Atlantic published an investigative report revealing that Meta faced an ethical dilemma while developing their AI model, Llama 3. The company opted to use pirated resources, notably Library Genesis (LibGen), which offers over 7.5 million books and 81 million research papers, due to the high cost and slow process of legal acquisition. This move, backed by Meta CEO Mark Zuckerberg, is now part of a copyright infringement lawsuit involving several authors.
5. <https://apnews.com/article/168b32059e70d0509b0a6ac407f37e8a> - In March 2025, French publishers and authors filed a lawsuit against Meta, accusing the company of using their copyrighted works without permission to train its artificial intelligence (AI) model. Three trade groups, including the National Publishing Union, the National Union of Authors and Composers, and the Societe des Gens de Lettres, collectively initiated the legal action in a Paris court. They allege that Meta has engaged in unauthorized use of their works to develop generative AI capabilities for its platforms, including Facebook, Instagram, and WhatsApp.
6. <https://www.ibpa-online.org/news/697191/IBPA-Statement-on-Metas-Use-of-Pirated-Books-to-Train-AI.htm> - In March 2025, the Independent Book Publishers Association (IBPA) released a statement supporting the Authors Guild in response to Meta's use of pirated books to train its AI model, Llama 3. The IBPA criticized Meta's decision to use pirated content instead of engaging in ethical licensing agreements with authors and publishers. The association emphasized the importance of authors' rights and called for fair compensation for the use of their works in AI development.
7. <https://arstechnica.com/tech-policy/2025/02/meta-torrented-over-81-7tb-of-pirated-books-to-train-ai-authors-say/> - In February 2025, Ars Technica reported on newly unsealed emails providing evidence that Meta torrented at least 81.7 terabytes of data from pirated book databases, including LibGen, to train its AI models. The authors' court filing alleged that Meta not only downloaded copyrighted material but also participated in 'seeding'—sharing the downloaded files with other users through torrent networks. This activity effectively made Meta 'a distributor of the very same pirated copyrighted material' used in its AI models.