# Meta’s AI memorisation of copyrighted books sparks billion-dollar legal threat



Recent revelations concerning artificial intelligence training practices have raised significant questions about copyright infringement, particularly regarding Meta's practices. A new study indicates that certain AI models, notably Meta's, have not only been trained on copyrighted books but have also memorised extensive portions of their content verbatim. This finding could have extensive legal ramifications, with potential implications for billions of dollars as ongoing cases unfold in the U.S. and the UK over the legality of these practices.

Authors and publishers have been increasingly vocal against AI companies using their works without permission for model training. In April, a coalition of authors and publishers voiced their concerns regarding Meta's usage of copyrighted materials, which they argue constitutes a violation of their rights. The core legal dispute centres around whether tech firms can legitimately use copyrighted materials without prior consent, with many AI developers claiming that their models generate entirely new compositions rather than direct replications of the original texts.

The new research, led by a team including Mark Lemley from Stanford University, highlights a significant discrepancy among various AI models. While most models do not recall the exact wording of texts from their training, Meta's Llama 3.1 model shows a concerning aptitude for reproducing sizeable segments of well-known works, including "Harry Potter and the Philosopher’s Stone" and "The Great Gatsby." The implications of these findings are severe: if a court determines that Meta's model has infringed on the copyright of even a small fraction of the Books3 dataset — a trove of nearly 200,000 titles used for training — damages could reach nearly $1 billion.

Despite evidence suggesting that Meta's AI practices may not align with fair use, company representatives assert that their approach is legitimate. Emil Vazquez, a spokesperson for Meta, contends that "fair use of copyrighted materials is vital" for the evolution of AI technologies. However, this stance has drawn scrutiny from legal experts, who assert that the core issue isn't merely about whether or not AI models can generate new works but hinges on whether they can legally draw from copyrighted content in the first place.

Further complicating matters, European legal action against Meta adds another layer to the discourse. In March 2025, prominent French authors and publishers accused Meta of "monumental pillage" for its unequitable use of protected works, drawing attention to the importance of adhering to copyright laws as outlined in the European Union's AI Act. Law firms in different jurisdictions are also weighing in, with Robert Lands highlighting that under UK law, the memorisation finding could lead to unfavourable results for AI firms, given the narrower exceptions available compared to the U.S. framework.

The ethical ramifications of utilising pirated materials for AI model training cannot be overlooked. A report indicates that Meta employees considered relying on piracy to acquire necessary training data due to the prohibitive costs of legal sourcing. This decision may resonate with broader challenges facing tech companies, all of whom risk undermining the original creative work and intellectual discussions that the very AI technologies aim to enhance.

As litigation continues and the technology evolves, the balance between innovation and intellectual property rights remains a contentious issue. Judges and legal experts are tasked with drawing clear boundaries, a complex challenge illustrated by Judge Vince Chhabria's observations during a recent court hearing, where he questioned the fairness of allowing companies to profit from unlicensed use of copyrighted material.

The outcome of these ongoing legal disputes may shape the landscape of AI development, potentially creating precedents that either facilitate or hinder the use of copyrighted materials for the training of generative models. As opinions clash and new evidence emerges, the contentious dialogue about copyright, creativity, and technology is far from settled.

### 📌 Reference Map:

* Paragraph 1 – [[1]](https://www.newscientist.com/article/2483352-metas-ai-memorised-books-verbatim-that-could-cost-it-billions/), [[5]](https://apnews.com/article/df4dec4aef8924d38d258212e0654a3d)
* Paragraph 2 – [[1]](https://www.newscientist.com/article/2483352-metas-ai-memorised-books-verbatim-that-could-cost-it-billions/), [[6]](https://www.reuters.com/legal/litigation/judge-meta-case-weighs-key-question-ai-copyright-lawsuits-2025-05-01/)
* Paragraph 3 – [[3]](https://www.reuters.com/technology/artificial-intelligence/french-publishers-authors-file-lawsuit-against-meta-ai-case-2025-03-12/), [[4]](https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/?utm_source=apple_news), [[2]](https://www.reuters.com/legal/litigation/google-fends-off-part-textbook-publisher-lawsuit-over-ads-2025-06-05/)
* Paragraph 4 – [[1]](https://www.newscientist.com/article/2483352-metas-ai-memorised-books-verbatim-that-could-cost-it-billions/), [[7]](https://apnews.com/article/168b32059e70d0509b0a6ac407f37e8a)
* Paragraph 5 – [[3]](https://www.reuters.com/technology/artificial-intelligence/french-publishers-authors-file-lawsuit-against-meta-ai-case-2025-03-12/), [[4]](https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/?utm_source=apple_news)
* Paragraph 6 – [[6]](https://www.reuters.com/legal/litigation/judge-meta-case-weighs-key-question-ai-copyright-lawsuits-2025-05-01/)

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

## Bibliography

1. <https://www.newscientist.com/article/2483352-metas-ai-memorised-books-verbatim-that-could-cost-it-billions/> - Please view link - unable to able to access data
2. <https://www.reuters.com/legal/litigation/google-fends-off-part-textbook-publisher-lawsuit-over-ads-2025-06-05/> - In June 2025, a Manhattan federal court partially dismissed a lawsuit against Google by leading educational publishers, including Cengage Learning, McGraw Hill, Macmillan Learning, and Elsevier. The publishers accused Google of unlawfully promoting pirated electronic versions of their textbooks through its search engine. U.S. District Judge Jennifer Rochon ruled that the publishers failed to substantiate claims of vicarious copyright infringement and violations of New York state law. However, she allowed their trademark infringement claim and a contributory copyright infringement claim to proceed. The publishers expressed satisfaction that key aspects of their complaint remain, criticizing Google for allegedly enabling piracy instead of protecting copyright holders. Google denied the allegations, stating that pirated e-books are not sold through its platform and that it does not profit from their sales. The case, Cengage Learning Inc. v. Google LLC, continues in the U.S. District Court for the Southern District of New York.
3. <https://www.reuters.com/technology/artificial-intelligence/french-publishers-authors-file-lawsuit-against-meta-ai-case-2025-03-12/> - In March 2025, major French publishing associations and authors filed a lawsuit against Meta, alleging the company used their copyrighted works without permission to train its artificial intelligence systems. The plaintiffs, including the National Publishing Union, the National Union of Authors and Composers, and the Society of People of Letters, accused Meta of copyright infringement and economic parasitism. They argued that Meta illegally used protected content to develop its AI models, describing it as a "monumental pillage." This lawsuit, filed in a Paris court, is part of a broader trend of similar legal actions against tech companies over the use of protected data for training generative AI systems.
4. <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/?utm_source=apple_news> - In March 2025, Meta faced an ethical dilemma while developing their AI model, Llama 3: should they pirate the vast amounts of high-quality writing needed for training due to the high cost and slow process of legal acquisition? They opted for pirated resources, notably Library Genesis (LibGen), which offers over 7.5 million books and 81 million research papers. This move, backed by Meta CEO Mark Zuckerberg, is now part of a copyright infringement lawsuit involving several authors. OpenAI has a similar history with using LibGen data. Meta employees discussed various strategies to mask and avoid legal repercussions while using pirated content. LibGen, created to provide accessible knowledge in resource-constrained regions, has faced multiple lawsuits, yet continues to thrive. Generative AI companies argue their models transform original work under "fair use," a defense still unresolved in court. Despite increasing knowledge accessibility, these practices risk devaluing original creative labor, hindering proper attribution, and impacting human intellectual debate. The larger question remains: how to manage and distribute knowledge that benefits society the most.
5. <https://apnews.com/article/df4dec4aef8924d38d258212e0654a3d> - In October 2024, Meta CEO Mark Zuckerberg was deposed in a lawsuit brought by authors, including Sarah Silverman, accusing Meta of using their copyrighted works without consent to train its AI. U.S. District Judge Thomas Hixson denied Meta's attempt to prevent Zuckerberg's deposition, citing evidence of his close involvement in AI initiatives. The class action, filed last year, claims Meta illegally downloaded digital copies of the authors' books. Prominent attorney David Boies has joined the case representing Silverman and other plaintiffs, which also includes noted authors and public figures. The suit is part of a wider set of legal actions in San Francisco and New York against AI developers like Anthropic, Microsoft, and OpenAI.
6. <https://www.reuters.com/legal/litigation/judge-meta-case-weighs-key-question-ai-copyright-lawsuits-2025-05-01/> - In May 2025, during a pivotal legal hearing in San Francisco, U.S. District Judge Vince Chhabria questioned Meta Platforms' legal justification for using copyrighted materials without permission to train its AI model, Llama. Authors including Junot Diaz and Sarah Silverman filed lawsuits in 2023, alleging that Meta used pirated versions of their books for AI training without compensation. The case focuses on whether such use constitutes "fair use" under U.S. copyright law, a crucial issue for the AI industry. Chhabria expressed skepticism about the fairness of allowing companies to use protected works to create potentially infinite competing outputs without licensing them, possibly undermining the market for the original work. Meta argued that their actions are transformative and fall under fair use, while plaintiffs contend it amounts to infringement that threatens their livelihoods. Though acknowledging the transformative nature of AI-generated content, Chhabria emphasized the need for plaintiffs to demonstrate specific market harm. The judge’s probing remarks underscored the complex balance courts must strike between fostering innovation and protecting intellectual property rights.
7. <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. The groups demand the removal of the data directories created by Meta for AI training. This legal move comes in the wake of the European Union's AI Act, which mandates compliance with copyright laws for AI systems. The case reflects ongoing tensions between the creative industries and tech companies over the use of data and copyrighted material.