Learned Publishing

LEARNED PUBLISHING Open Access

OPINION OPEN ACCESS

Science Behind a Paywall: Restricted Access Limits the Promise of Artificial Intelligence

Haoyi Zheng[1] | Huichun Zhan[2,3]

[1]Saint Francis Hospital, The Heart Center, Roslyn, New York, USA | [2]Department of Medicine, Stony Brook School of Medicine, Stony Brook, New York, USA | [3]Medical Service, Northport VA Medical Center, Northport, New York, USA

Correspondence: Haoyi Zheng (haoyi.zheng@chsli.org)

Received: 13 December 2025 | **Revised:** 13 March 2026 | **Accepted:** 20 April 2026

1 | Introduction

Artificial intelligence (AI) is increasingly integrated into scientific work. Large language models (LLMs) like Grok, ChatGPT, Claude, DeepSeek, and Gemini can summarise literature, suggest hypotheses, revise manuscripts, and even assist with peer review (Mishra et al. 2024). These systems rely on access to vast corpora of scientific literature and other human-generated data to identify patterns, maintain context, and generate content. However, there are growing concerns about the accuracy of AI-generated content and LLM's poor performance in scientific research and reasoning (McCoy et al. 2025; Kalai et al. 2025; Thomas et al. 2026). Like human researchers or clinicians, LLMs must "read" broadly and deeply to learn. Yet, unlike humans with institutional or library subscriptions to full-text articles, AI systems face a critical and often overlooked barrier: paywall.

2 | Paywalls Restrict AI Access to Scientific Knowledge

A paywall is a digital barrier set by a publisher that blocks full-text journal articles unless a subscription or payment is made. The proportion of paywalled literature has gradually declined with the expansion of open-access publishing. However, recent reports (2024–2025) estimate that approximately 50% of scholarly articles remain behind paywalls, and the growth of open access appears to have plateaued in recent years (Delta Think 2025). In biomedicine, PubMed currently indexes over 39 million biomedical research articles, of which only about 14 million were open-access as of March 2026 (PubMed 2026; Free Full Text[sb] - Search Results - PubMed 2026). This suggests that 64% of biomedical articles remain paywalled and unavailable to AI for training or analysis. Major publishers such as Elsevier, Springer Nature, Wiley, Taylor & Francis, and Sage control large portions of the academic literature and use licences that restrict automated access and roughly 50% of global research articles published by these five firms (Lucy Montgomery ECB and Huang 2025).

Paywalls secure publishers' revenue by charging for full-text access. Much of their revenue comes from the hefty fees paid by academics and libraries. Additional income comes from author-paid open-access fees. Several major publishers have been criticised for profit margins exceeding 30%, while relying heavily on unpaid scholar labour (authors, peer reviewers, and academic editors) (Trueblood et al. 2025).

Without authorization, AI systems cannot read paywalled articles, including landmark studies. As a result, LLMs are trained mostly on abstracts, media coverage, preprints, open-access publications, and public databases, leaving much of the peer-reviewed articles out of reach. These sources are valuable but often lack the depth and precision required for AI to perform scientific reasoning. Abstracts provide only skeletal methods, results, and conclusions; media summaries are often superficial or fragmented. For general summaries or conclusions, abstract or news coverage may suffice; but for in-depth data and detailed methodology analysis, full-text articles including supplemental materials are required. Moreover, major media outlets are increasingly paywalled their content. Notably, the New York Times has sued OpenAI over unauthorised use of paywalled articles (Grynbaum and Mac 2023). Beyond direct conflict with the publishers' paywalls, another notable ethical and legal concern in developing major LLMs is the potential use of illegally

Key Points

- Paywalls restrict AI access to approximately 50% of full-text scholarly articles, limiting training on high-quality data essential for scientific accuracy.
- Lack of full-text access for AI risks a self-reinforcing cycle of low-quality outputs that potentially erode the integrity of scholarly publishing.
- Paywalls fund vital publishing infrastructure and are legally protected, making paywall removal unsustainable.
- Emerging licencing partnerships together with RAG offer practical ways to grant AI secure access while sustaining publishers.
- Urgent action to bridge this gap is needed before AI-generated synthetic data dominates.

obtained full-text materials including pirated scholarly articles, books, or datasets from shadow libraries like Library Genesis (LibGen) or Sci-Hub. Meta, the developer of Llama models, was sued for using pirated LibGen datasets for model training (Knibbs 2025); Anthropic was sued for downloading large volume of pirated books from LibGen and Pirate Library Mirror to train its claude (The Associated Press 2025). These cases have raised infringement claims related to the acquisition and retention, and use of unlawfully acquired copyrighted works in AI development.

As with any fine art, the essence of scientific endeavour lies in its attention to detail; no masterpiece, in science or art, is possible without it. Lack of access to full-text publications is a major limitation on the accuracy of computational text-mining and AI systems (Westergaard et al. 2018). Full access to publications includes tables, figures, methods, detailed discussions, and supplemental data such as protocol and additional datasets. AI models trained on incomplete datasets are more likely to produce inaccurate summaries, overgeneralizations, or fabricated or misleading information (hallucination) (Zheng and Zhan 2023). Even when grammatically sound and structurally coherent, AI-generated content often lacks scientific rigour or fidelity to evidence that human experts demand. No matter how advanced AI becomes, its performance will always depend on the quality and completeness of the data it consumes. Just as human researchers rely on solid evidence or data to form sound conclusions, AI too requires full or meaningful access to the scientific literature to perform reliably.

3 | The Feedback Loop of Exclusivity: Publishers and AI Can Produce Mediocrity

The credibility of scientific publishing is already challenged by the proliferation of low-quality publications, driven by academic incentives to publish and commercial publishing models. AI is amplifying this problem. The rise of AI-generated content that is written by AI, reviewed by AI, and read by AI could create a self-reinforcing cycle without human scrutiny. When AI models are trained primarily on fragmented or limited data, they are more likely to produce shallow and generic summaries, unsubstantiated statements, and hallucinations at a mass scale. Such flawed output, if recycled into AI training, may perpetuate a downward spiral of misinformation, mediocrity, and degradation. A study showed that constant learning from AI-generated content caused models to degrade their performance, eventually leading to a complete model collapse. In practical terms, model collapse occurs when AI-generated outputs are recycled as training data, gradually eroding or even replacing the original human-generated content over time (Shumailov et al. 2024). The risk is particularly relevant for scientific applications, where reliable interpretation depends on precise methodological details and accurate results. If AI systems are trained primarily on summaries or AI-generated interpretations rather than original scientific full-text publications, the fidelity of scientific knowledge embedded in these models may progressively deteriorate. An AI-driven "infodemic" is already underway (De Angelis et al. 2023). Ironically, the paywalls designed to protect the value of literature and generate profit may accelerate this decline by denying AI access. In an era when AI could advance evidence synthesis and democratise scientific knowledge, paywall has become more than a monetary barrier and it now poses a threat not only to the integrity of scientific publishing, but also to science itself.

Tensions between AI and publishers are not new. In 2018, when Springer Nature announced the launch of Nature Machine Intelligence, a subscription-based journal, it faced swift backlash from the AI community. Its closed-access model clashed with widespread calls for open data sharing in AI research. In protest, thousands of researchers and students, including many leaders in the field, signed a public pledge to boycott the journal, vowing not to submit, review, or serve on its editorial board (Lawrence 2018).

Meanwhile, many leading AI platforms, such as GPT-4, Claude, and Gemini, now charge subscription fees for access to their advanced features. This raises a parallel question: should AI trained on open-access data or publicly funded research be commercialised? The monetization of AI outputs built on freely shared knowledge mirrors longstanding concerns in academic publishing, where public contributions and unpaid academic labour are commodified.

4 | Paywalls Sustain Scientific Publishing Ecosystem and Are Protected by the DMCA

The long-standing presence for paywalls has defensible rationales. Paywalls finance the editorial and community infrastructure that makes scientific publications trustworthy. Subscription revenue (alongside advertising and licenced reuse) funds manuscript triage, professional editing and production, integrity checks, long-term archiving, and platform upkeep. For many important scientific publications, producing companion blogs, videos, and podcasts all entails labour and cost. The income also supports fellowships, meetings, education, travel grants, and advocacy. Eliminating paywalls without alternative revenue stream would jeopardise these community missions and the quality controls.

Furthermore, paywalls as a digital access control are protected under US law. The Digital Millennium Copyright Act

(DMCA) (U.S. Copyright Office 2020) is a US copyright law that addresses copyright issues in the digital age. DMCA prohibits bypassing measures that control access to copyrighted works without authorization. However, the legal framework governing AI access to copyrighted materials remains largely unsettled. The US Copyright Office's 2025 report, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (U.S. Copyright Office 2025), provides the comprehensive analysis to date, and acknowledge that the legality of large-scale training on copyrighted works is highly context-dependent and remains unresolved in many situations. The key question, as the report indicated, is whether acts that may constitute prima facie infringement can be justified under the fair use doctrine.

Proponents of AI training argue that using copyrighted texts to train models may constitute transformative fair use, drawing parallels to precedents such as *Authors Guild* versus *Google* and *Authors Guild* versus *HathiTrust* (West Publishing Co 2015). In those cases, large-scale digitization and text mining of copyrighted books were deemed lawful because the resulting systems enabled new forms of search and analysis rather than replacing the original works. Critics, however, argue that generative AI systems may compete with or substitute for the original publications, potentially undermining the market for those works. This issue is currently being examined in several ongoing legal disputes involving publishers and AI developers (U.S. Copyright Office 2025).

Therefore, paywalls cannot simply taken as approaches against openness or as profit-driven businesses only; they function as a financing and copyright managing tool that preserves editorial rigour, sustains society missions, and mitigates inequities created by author-paid article processing charges under open access. At the same time, AI's rapid development poses an unprecedent challenge to established copyright frameworks. To ensure that AI development respect core legal and policy principles under the DMCA, including authorship and originality, has become imperative.

5 | Potential Solutions: Build Publishers-AI Partnerships

To ensure AI learning over time, we must preserve access to the original, human-generated data and ensure a continued supply of data not generated by LLMs (Shumailov et al. 2024). However, it would be unrealistic to expect publishers to drop paywalls and grant unrestricted access to AI developers without meaningful financial incentives. Open access has been a leading reform, but in practice, it shifts costs from readers to authors via substantial article processing charges. It has failed to halt the commercialization of publications (Trueblood et al. 2025). Furthermore, the open access model has been exploited by predatory journals that publish almost anything for profit. At the same time, paper mills sell fabricated or fake manuscripts at an industrial scale (Lawrence 2018).

The unsettled legal frameworks highlight the continued uncertainty surrounding AI training on copyrighted materials and reinforce the importance of licenced partnerships and secure-access mechanisms as practical near-term solutions. A practical path forward is collaboration grounded in long-term sustainability. One model is to have formal licencing agreements between publishers and AI developers, similar to the fee-based subscription used by academic libraries. Agreements should include data security, confidentiality, and proper attribution, and allow AI models to process content inside a secure, publisher-controlled environment so training occurs without copying or redistributing the texts. Another approach can be drawn from the biopharmaceutical industry (Mock et al. 2023). Drug companies use federated and active learning consortia that allow participants to collaboratively train AI models on proprietary data while keeping sensitive information secure. This structure preserves data privacy while maximising data sharing. A similar framework could be adopted for partnerships between publishers and AI developers, facilitating access to high-quality content while safeguarding intellectual property. Encouragingly, such partnerships have already begun. In May, 2024, Taylor & Francis signed a non-exclusive Partnership and Data Access Agreement with Microsoft for access to its nearly 3000 academic journals (Informa 2024). In July 2025, Wiley announced partnership with Anthropic that allows the Claude AI system to access the publisher's licenced content (Wiley 2025).

Furthermore, rapid advances in AI have made such partnerships between AI developers and publishers increasingly feasible and effective. Unlike traditional LLM models, which rely primarily on information embedded during their pretraining (Gupta et al. 2024), emerging retrieval-augmented generation (RAG) provides a highly effective framework for collaboration between AI developers and publishers. First introduced by Lewis et al. in 2020 (Lewis et al. 2020), RAG combines a non-parametric retriever with a parametric generative model, which substantially enhances the capabilities of LLMs. This combination enables AI models not only to generate fluent text but also to ground their outputs in real-time, up-to-date data (Gupta et al. 2024). Importantly, RAG allows AI systems to retrieve relevant content at the time of query rather than embedding paywalled material directly into training datasets. Information can instead be accessed through existing institutional subscriptions, licenced publisher application programming interfaces (APIs), or other secure access platforms (Gupta et al. 2024). In such a system, publishers retain control over their copyrighted material while AI systems can analyse and synthesise information when permitted by licencing agreements or subscriptions. By separating model training from information retrieval, RAG-based approaches help balance the interests of publishers, researchers, and AI developers while also improving the accuracy and reliability of AI-generated scientific contents. Since 2024, RAG has become a core component of many leading LLM systems.

At the same time, preprint repositories such as arxiv, biorxiv, and medrxiv have served as valuable sources of data for AI training. These preprint repositories have played a pivotal role in the rapid dissemination and communication of emerging research findings, enabling rapid sharing of results before formal peer review and journal publication. They represent a substantial and growing portion of openly accessible scientific literature. However, preprints lack peer review, a cornerstone of scientific publishing, and their trustworthiness remains a concern. Thus, preprint repositories cannot substitute for access to the peer-reviewed, rigorously scrutinised literature behind paywalls.

Today, on one hand, it is in the public interest for AI models to be trained on high-quality publications; on the other, both publishers and AI developers are creating information asymmetries and monopolies of knowledge, and are increasingly treating scientific knowledge as a commodity. They operate within exclusive, profit-driven systems that determine who may access or benefit from scientific knowledge. Meanwhile, unpaid scholar labour of the authors, peer reviewers, and many academic editors underpins the publishing enterprise. How to balance financial incentives with openness remains unresolved. Without reform, the convergence of these exclusive models risks creating a self-reinforcing cycle of decline and eroding the science's credibility. Such an outcome would have profound and unacceptable consequences for both science and society.

6 | Summary

AI holds potential to enhance human creativity and accelerate discovery, but its performance depends on complete and high-quality data. Paywalls remain a practical barrier. Both commercial publishing and AI developers currently operate within exclusive, profit-driven systems. Without reform, AI development risks a self-reinforcing decline in quality, and at the same time erodes the credibility of science. The solution is not to dismantle paywalls, but to reimagine the system through publisher-AI collaboration. It is predicted that human-produced data will ultimately be exhausted for AI training and synthetic data will come to dominate. It is urgent to train models now on comprehensive, human-generated corpora. Only by unlocking the literature for AI can we fully unlock AI for the literature, and for science.

Author Contributions

Haoyi Zheng prepared the first draft; Huichun Zhan provided key revisions and rewriting.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

De Angelis, L., F. Baglivo, G. Arzilli, et al. 2023. "ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health." *Frontiers in Public Health* 11: 1166120. https://doi.org/10.3389/fpubh.2023.1166120.

Delta Think. 2025. "News & Views: Market Sizing Update 2025 – Has OA Recovered Its Mojo?." https://www.deltathink.com/news-views-market-sizing-update-2025-has-oa-recovered-its-mojo/.

Free Full Text[sb] - Search Results - PubMed. 2026. "National Library of Medicine (U.S.)." https://pubmed.ncbi.nlm.nih.gov/?term=free+full+text%5Bsb%5D.

Grynbaum, M. M., and R. Mac. 2023. "The Times Sues OpenAI and Microsoft Over AI Use of Copyrighted Work." *New York Times*, December 27, 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Gupta, S., R. Ranjan, and S. N. Singh. "A Comprehensive Survey of Retrieval-Augmented Generation (Rag): Evolution, Current Landscape and Future Directions." arXiv Preprint arXiv:241012837. (2024).

Informa. 2024. "AI Partnership and Data Access Agreement." https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc---market-update.pdf.

Kalai, A. T., O. Nachum, S. S. Vempala, and E. Zhang. 2025. "Why Language Models Hallucinate." arXiv Preprint arXiv:250904664.

Knibbs, K. 2025. "Kadrey v. Meta Platforms: Allegations of Meta's Use of Pirated LibGen Data for Llama Training." *Wired*, January 9, 2025. https://www.wired.com/story/new-documents-unredacted-meta-copyright-ai-lawsuit/.

Lawrence, N. 2018. "Why Thousands of AI Researchers Are Boycotting the New Nature Journal." https://www.theguardian.com/science/blog/2018/may/29/why-thousands-of-ai-researchers-are-boycotting-the-new-nature-journal.

Lewis, P., E. Perez, A. Piktus, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive nlp Tasks." *Advances in Neural Information Processing Systems* 33: 9459–9474.

Lucy Montgomery ECB, and K. Huang. 2025. "Academic Publishing Is a Multibillion-Dollar Industry. It's Not Always Good for Science."

McCoy, L. G., R. Swamy, N. Sagar, et al. 2025. "Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study." *NEJM AI* 2: AIdbp2500120. https://doi.org/10.1056/AIdbp2500120.

Mishra, T., E. Sutanto, R. Rossanti, et al. 2024. "Use of Large Language Models as Artificial Intelligence Tools in Academic Research and Publishing Among Global Clinical Researchers." *Scientific Reports* 14: 31672. https://doi.org/10.1038/s41598-024-81370-6.

Mock, M., S. Edavettal, C. Langmead, and A. Russell. 2023. "AI Can Help to Speed Up Drug Discovery - But Only If We Give It the Right Data." *Nature* 621: 467–470. https://doi.org/10.1038/d41586-023-02896-9.

PubMed. 2026. "National Library of Medicine (U.S.)." https://pubmed.ncbi.nlm.nih.gov/.

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631: 755–759.

The Associated Press. 2025. "Anthropic v. Bartz: Class-Action Over Anthropic's Downloading of Pirated Books From LibGen/PiLiMi for Claude Training." *NPR*, September 5, 2025. https://www.npr.org/2025/09/05/g-s1-87367/anthropic-authors-settlement-pirated-chatbot-training-material.

Thomas, L. D. W., A. K. G. Romasanta, and P. L. Pujol. 2026. "Jagged Competencies: Measuring the Reliability of Generative AI in Academic Research." *Journal of Business Research* 203: 115804. https://doi.org/10.1016/j.jbusres.2025.115804.

Trueblood, J. S., D. B. Allison, S. M. Field, et al. 2025. "The Misalignment of Incentives in Academic Publishing and Implications for Journal Reform." *Proceedings of the National Academy of Sciences* 122: e2401231121. https://doi.org/10.1073/pnas.2401231121.

U.S. Copyright Office. 2020. "The Digital Millennium Copyright Act." https://www.copyright.gov/dmca.

U.S. Copyright Office. 2025. "Copyright and Artificial Intelligence, Part 3: Generative AI Training." https://www.copyright.gov/ai/?utm_source=chatgpt.com.

West Publishing Co. 2015. "Authors Guild, Inc. v. Google, Inc." https://www.copyright.gov/fair-use/summaries/authorsguild-google-2dcir2015.pdf.

Westergaard, D., H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak. 2018. "A Comprehensive and Quantitative Comparison of Text-Mining in 15 Million Full-Text Articles Versus Their Corresponding Abstracts." *PLoS Computational Biology* 14: e1005962. https://doi.org/10.1371/journal.pcbi.1005962.

Wiley. 2025. "Wiley Partners With Anthropic to Accelerate Responsible AI Integration Across Scholarly Research." https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-Partners-with-Anthropic-to-Accelerate-Responsible-AI-Integration-Across-Scholarly-Research/default.aspx.

Zheng, H., and H. Zhan. 2023. "ChatGPT in Scientific Writing: A Cautionary Tale." *American Journal of Medicine* 136, no. 725–6: e6. https://doi.org/10.1016/j.amjmed.2023.02.011.