The Paradigm Shift in Health Sciences Literature: Charting the Future of Large Language Models in Scientific Publishing

AI Author: Gemini, 1.5 Pro.

Developer: Google DeepMind

Role of AI Author: Literature synthesis, structural organization, and manuscript drafting

Human Prompter: Diego A. Forero, MD, PhD.

School of Health and Sport Sciences, Fundación Universitaria del Área Andina, Bogotá. Colombia. Email: dforero41@areandina.edu.co

Abstract

The exponential expansion of biomedical literature, coupled with the growing demand for rapid clinical knowledge dissemination, has created an unsustainably high workload for researchers, peer reviewers, and journal editors. Large Language Models (LLMs) represent a transformative architectural pivot capable of automating, optimizing, and reshaping workflows across the health sciences publishing lifecycle. This paper comprehensively analyzes the current and future applications of LLMs in manuscript preparation, peer review workflows, and editorial curation. Furthermore, we examine the profound technical and ethical vulnerabilities inherent to generative artificial intelligence in medicine—including factual confabulations, demographic biases, and data privacy conflicts with regulatory frameworks like HIPAA and GDPR. Finally, we evaluate the emerging consensus frameworks, such as the FUTURE-AI guidelines and updated International Committee of Medical Journal Editors (ICMJE) criteria, advocating for a hybrid human-AI symbiotic model that preserves scientific integrity while maximizing technological efficacy.

Keywords: Large Language Models, Scientific Publishing, Health Sciences, Peer Review, Publication Ethics, FUTURE-AI, Bioethics.

1. Introduction

Scientific publishing within the health sciences operates under a dual imperative: it must maintain uncompromising empirical rigor while accelerating the dissemination of clinical discoveries to optimize patient care. However, the contemporary biomedical research ecosystem is facing unprecedented strain. The volume of publications, clinical trials, and systematic reviews has grown exponentially, outpacing the capacity of the traditional peer review model and precipitating cognitive overload among medical professionals (Ahn, 2024; Zhang et al., 2025).

The introduction of transformer-based Large Language Models (LLMs)—characterized by billions of parameters trained on vast, multimodal text corpora—has emerged as a disruptive force capable of fundamentally shifting this paradigm. By analyzing statistical patterns over sequences of text, LLMs possess an unparalleled capability to comprehend, summarize, generate, and critique complex scientific prose (Schrager et al., 2025; Telenti et al., 2024). In the health sciences, where data modalities extend beyond standard prose to encompass electronic medical records, genomic sequences, and clinical imagery, LLMs offer a unified cognitive layer capable of bridging raw laboratory discovery with polished academic reporting (Telenti et al., 2024).

This paper provides a critical exploration of the future of LLMs within health sciences academic publishing. It explores how these tools can democratize scientific communication, outlines the operational risks they introduce to the scholarly record, and frames the governance mechanisms necessary to preserve the sanctity of evidence-based medicine.

2. Applications of LLMs in Medical Manuscript Preparation

The initial phase of the scholarly lifecycle—conceptualization, literature synthesis, and drafting—is incredibly labor-intensive. LLMs are rapidly shifting from passive word-processing assistants to proactive research co-pilots across several operational dimensions.

2.1 Literature Retrieval and Knowledge Synthesis

Traditional database queries often rely heavily on rigid Boolean strings, which can inadvertently exclude relevant clinical insights due to semantic variations. LLMs, particularly when paired with semantic search architectures and Retrieval-Augmented Generation (RAG), allow researchers to engage in conversational synthesis over vast text databases (Gencer & Gencer, 2025; Zhang et al., 2025). These systems can extract granular details from thousands of disparate publications, map disease-definition variations, and summarize treatment methodologies at a scale impossible for human investigators acting alone (Gencer & Gencer, 2025).

2.2 Democratizing Global Scientific Communication

A persistent structural inequality in health sciences publishing is the linguistic barrier faced by non-native English-speaking clinicians and researchers. Manuscripts containing high-quality clinical data are frequently rejected by premium journals due to stylistic inconsistencies or grammatical oversights. LLMs address this disparity by serving as highly advanced language editors, stylistic calibrators, and real-time translators (Ahn, 2024). By smoothing syntax, refining vocabulary, and standardizing structural elements, LLMs allow global researchers to present their empirical findings equitably, shifting the editorial focus back to scientific merit rather than linguistic fluency.

2.3 Multimodal Integration and Formatting

The contemporary health sciences landscape requires the synthesis of heterogeneous data types. Advanced LLMs excel at processing multimodal inputs, enabling the simultaneous analysis of clinical metadata, protein sequences, and chemical structures alongside standard textual outputs (Telenti et al., 2024; Zhang et al., 2025). In manuscript preparation, these models can automate the creation of data tables, generate baseline code for statistical validation, and align reference citations with target journal guidelines, saving significant manual labor (Ahn, 2024).

3. LLMs in the Peer Review and Editorial Ecosystem

The peer review system is facing a systemic bottleneck, driven by a shortage of qualified human reviewers relative to the sheer volume of manuscript submissions. LLMs offer a potential solution to this crisis by streamlining editorial workflows.

3.1 Manuscript Screening and Triage

Upon submission, manuscripts undergo a preliminary editorial screening to ensure compliance with formatting rules, ethical disclosures, and core stylistic requirements. Editors are increasingly deploying specialized LLMs to automate this initial triage phase (Ahn, 2024). These models can rapidly detect structural errors, identify plagiarism, and flag basic methodology deficits before human editors review the text.

LLM Application Phase	**Primary Mechanisms / Capabilities**	**Major Limitations & Vulnerabilities**
Pre-Submission / Writing	Language editing, automated translation, referencing alignments, structural smoothing.	Reference fabrication, amplification of underlying author bias, loss of distinct academic voice.
Editorial Screening	Automated plagiarism checking, format validation,	Inability to evaluate deep intellectual novelty; risk

LLM Application Phase	**Primary Mechanisms / Capabilities**	**Major Limitations & Vulnerabilities**
	methodology completeness grading.	of high false-positive rejection rates.
Peer Review Evaluation	Structural error detection, baseline validation check, statistical coding replication.	"Black box" reasoning, systemic leniency (overly positive reviews), data privacy violations.

3.2 Automated Reviewer Comments and Hybrid Workflows

Empirical studies demonstrate that LLMs can generate peer review feedback that significantly overlaps with the comments provided by human experts (Ahn, 2024). LLMs can identify mathematical discrepancies, flag missing control groups in clinical designs, and assess the risk of bias within systematic reviews (Ahn, 2024). However, current iterations demonstrate distinct behavioral biases. For instance, LLMs frequently generate overly positive evaluations and lack the deep domain experience required to gauge the conceptual novelty or potential real-world clinical impact of a study (Ahn, 2024).

Consequently, the future of peer review is not fully autonomous, but rather a hybrid framework. In this model, AI tools handle lower-order reviewing duties—such as proofreading, compliance checks, and structural verification—allowing human experts to focus on higher-order evaluative tasks like conceptual validity, ethical feasibility, and clinical clinical relevance (Ahn, 2024).

4. Ethical, Technical, and Legal Vulnerabilities

While the benefits of LLMs are substantial, their integration into health sciences publishing introduces complex risks. Because medical research directly influences clinical behavior and patient outcomes, errors within published medical literature can have serious, real-world safety implications.

4.1 Factual Confabulations and Hallucinations

A core characteristic of LLMs is their probabilistic nature; they generate sequences of text based on statistical likelihood rather than an underlying understanding of absolute factual truth (Schrager et al., 2025). This architectural trait leads to "hallucinations"—the generation of highly plausible but entirely fabricated information (Ahn, 2024; Schrager et al., 2025). In medical writing, this can manifest as fake patient data, fabricated biochemical interactions, or entirely invented journal references (Ahn, 2024).

Even advanced multimodal models, such as GPT-4V, exhibit this vulnerability. When evaluated on clinical image challenges, these models may identify the correct diagnosis while providing deeply flawed or entirely fabricated rationales for their choice (Ahn, 2024). The dissemination of such errors within the peer-reviewed record poses a direct risk to clinical safety.

4.2 The "Paywall Blind Spot" and Information Bias

The empirical reliability of an LLM is inherently restricted by the scope and quality of its training data. A significant limitation for academic publishing tools is the "paywall blind spot." While public open-access repositories like PubMed Central and arXiv are widely represented in major training datasets, premium subscription content from major publishers (e.g., Elsevier, Springer Nature, Wiley) is often blocked by authentication barriers (Ahn, 2024).

As a result, LLMs are frequently trained on text-scraped data where open-access publications are overrepresented, while paywalled, high-impact clinical trials and rigorous methodological breakdowns are underrepresented. This creates a systemic information bias. Authors relying solely on LLMs for literature synthesis risk missing critical, paywalled methodological details, which can compromise the comprehensiveness of their research.

4.3 Demographic Bias Preservation

LLMs tend to mirror and amplify the systematic demographic biases present within their training data. In the medical domain, this risk is particularly acute, as models can propagate outdated, race-based medical assumptions, false characterizations of pain thresholds, or biased diagnostic metrics across diverse populations (Ahn, 2024). If authors rely uncritically on LLMs to draft clinical discussions or synthesize public health strategies, they risk institutionalizing these demographic biases within the peer-reviewed literature, potentially exacerbating health disparities.

4.4 Data Privacy, Confidentiality, and Regulatory Compliance

The peer review process is built on strict confidentiality. When a reviewer uploads an unpublished manuscript into a commercial, cloud-hosted LLM platform to generate summary text or feedback, that intellectual property may be ingested to train future iterations of the model (Ahn, 2024). This constitutes a serious breach of confidentiality and a violation of intellectual property rights (Ahn, 2024).

Furthermore, integrating clinical data or unstructured electronic health records into external generative AI platforms raises significant legal issues under privacy frameworks like HIPAA and GDPR (Ahn, 2024). Developing secure, locally deployed, open-source LLM workflows that comply with these strict regulatory frameworks remains a significant technical challenge for research institutions (Ahn, 2024).

5. Policy Frameworks and Consensus Guidelines

To safeguard the scientific record, publishers, editorial associations, and international AI consortia have established strict ethical guidelines governing the deployment of generative AI.

5.1 Authorship and Accountability Denials

The consensus across leading editorial bodies—including the Committee on Publication Ethics (COPE), the World Association of Medical Editors (WAME), and the International Committee of Medical Journal Editors (ICMJE)—is unambiguous: **Large Language Models cannot be listed as authors on scientific publications** (Ahn, 2024).

Authorship carries strict intellectual accountability for the accuracy, integrity, and ethical compliance of the work. Because AI tools cannot take legal or moral responsibility for their outputs, they do not satisfy these criteria (Ahn, 2024).

ICMJE Core Principle Summary: Authors bear full, unshared responsibility for verifying all content generated or assisted by artificial intelligence tools. Any inclusion of erroneous data or fabricated references constitutes scientific misconduct on the part of the human authors (Ahn, 2024; Ganjavi et al., 2024).

5.2 Mandatory Disclosure and Transparency Metrics

Current editorial standards require complete transparency regarding AI usage. Under modern guidelines, such as those adopted by the *Annals of Geriatric Medicine and Research*, authors must declare the deployment of generative AI tools both within their cover letters and in a dedicated section of the manuscript, specifying the exact model name, version number, manufacturer, and scope of application (Ahn, 2024). While routine language editing or spell-checking typically does not require formal disclosure, any substantive contribution to data analysis, literature synthesis, or text generation must be explicitly detailed (Ahn, 2024).

5.3 The FUTURE-AI Consortium Framework

To move beyond ad-hoc journal policies, the international FUTURE-AI Consortium established a cohesive, multi-disciplinary framework for the responsible deployment of AI in health sciences (Lekadir et al., 2025). The guideline is organized around six core principles:

1. **Fairness:** Ensuring AI tools perform consistently across diverse demographic subgroups, actively identifying and minimizing underlying training bias (Lekadir et al., 2025).
2. **Universality:** Standardizing models to ensure operational utility across varying clinical environments and technical infrastructures (Lekadir et al., 2025).
3. **Traceability:** Documenting the entire lifecycle of the AI tool—including data curation, exact prompt structures, optimization details, and stochasticity handling (Lekadir et al., 2025).
4. **Usability:** Ensuring human-centric designs that allow clinicians and reviewers to interpret outputs easily without excessive technical training (Lekadir et al., 2025).
5. **Robustness:** Validating models against unexpected data shifts, technical variations, and adversarial inputs to prevent clinical errors (Lekadir et al., 2025).
6. **Explainability:** Developing interpretable models that explicitly detail the clinical or statistical rationale behind an output, moving away from uninterpretable "black box" systems (Lekadir et al., 2025).

6. Future Horizons: The Next Decade of AI-Symbiotic Publishing

Looking toward the next decade, the role of LLMs in health sciences publishing will likely transition from basic assistive automation to a fully integrated, collaborative workflow.

6.1 Real-Time Peer Review and Dynamic Updating

The traditional, static "publish-and-forget" format of medical journals is increasingly out of step with the rapid pace of clinical data generation. Future publishing paradigms may leverage specialized, locally hosted medical LLMs to provide continuous, real-time post-publication peer review. As new clinical trials or epidemiological data emerge, automated AI systems could screen existing publications, dynamically updating systematic reviews, meta-analyses, and clinical guidelines while highlighting contradictions or confirming medical hypotheses as new evidence accumulates.

6.2 Shift in Academic Valuations and the "Human Core"

As LLMs become highly proficient at generating grammatically pristine, stylistically polished scientific prose, the historical correlation between fluid writing and high-quality science will decouple (Ahn, 2024). Editorial boards and peer reviewers will adapt by placing less emphasis on prose mechanics, focusing instead on the core human elements of research: the prospective design of clinical trials, the execution of laboratory experiments, ethical oversight, and the nuanced interpretation of anomalous results (Ahn, 2024). The value of a scientific manuscript will reside in its empirical integrity and conceptual novelty, rather than its stylistic execution.

7. Conclusion

Large Language Models present a powerful opportunity to optimize health sciences publishing, offering tools to mitigate reviewer burnout, democratize global scientific writing, and synthesize vast amounts of clinical data. However, their integration introduces significant risks to scientific integrity, including factual hallucinations, data privacy vulnerabilities, and demographic biases.

To navigate this transition safely, the medical research community must reject both uncritical adoption and outright resistance. The future of health sciences publishing lies in a collaborative, human-AI symbiotic workflow. By establishing rigorous governance frameworks—such as the FUTURE-AI guidelines—and maintaining absolute human accountability for the verification of empirical data, the scientific community can leverage generative AI to accelerate discovery while safeguarding the evidence-based foundation of clinical medicine.

References

- Ahn, S. (2024). The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. *The Korean Journal of Physiology & Pharmacology*, *28*(5), 393–401. https://doi.org/10.4196/kjpp.2024.28.5.393
- Ganjavi, C., Eppler, M B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G. S., Gill, I. S., & Cacciamani, G. E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *BMJ*, *384*, e077192. https://doi.org/10.1136/bmj-2023-077192
- Gencer, G., & Gencer, K. (2025). Large Language Models in Healthcare: A Bibliometric Analysis and Examination of Research Trends. *Journal of Multidisciplinary Healthcare*, *Volume 18*, 223–238. https://doi.org/10.2147/jmdh.s502351
- Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., Amugongo, L. M., Fromont, L. A., Maier-Hein, L., Cerdá-Alberich, L., Martí-Bonmatí, L., & Cardoso, M. J. (2025). FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, e081554. https://doi.org/10.1136/bmj-2024-081554
- Schrager, S., Seehusen, D. A., Sexton, S., Richardson, C. R., Neher, J., Pimlott, N., Bowman, M. A., Rodríguez, J., Morley, C. P., Li, L., & Dera, J. D. (2025). Use of AI in Family Medicine Publications: A Joint Editorial From Journal Editors. *The Annals of Family Medicine*, *23*(1), 1–4. https://doi.org/10.1370/afm.240575
- Telenti, A., Auli, M., Hie, B. L., Maher, C., Saria, S., & Ioannidis, J. P. A. (2024). Large language models for science and medicine. *European Journal of Clinical Investigation*, *54*(1). https://doi.org/10.1111/eci.14183
- Zhang, K., Meng, X., Yan, X., Ji, J., Liu, J., Xu, H., Zhang, H., Liu, D., Wang, J., Wang, X., Gao, J., Wang, Y., Shao, C., Wang, W., Li, J., Zheng, M., Yang, Y., & Tang, Y. (2025). Revolutionizing Health Care: The Transformative Impact of Large Language Models in Medicine. *Journal of Medical Internet Research*, *27*, e59069. https://doi.org/10.2196/59069