IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 25, Issue 5 Ser. 2 (May. 2026), PP 62-65
www.iosrjournals.org

The Role Of Prompt Engineering In Optimizing The Quality And Reliability Of Generative AI For Medical Writing: A Narrative Review

Mitesh Mohan Hood
Senior Manager – Medical Writing, WPP Production

Abstract
Introduction: *The rapid evolution of generative artificial intelligence (GenAI) is reshaping the methodological landscape of medical research. Tools such as ChatGPT, Claude, and Gemini are being integrated into all stages of the research lifecycle, from hypothesis generation to peer review. However, the quality and reliability of these powerful tools are highly dependent on user interaction.*
Methods: *A targeted literature search was conducted in the PubMed database for articles published between 2024 and 2026 using a specific search string combining MeSH terms and keywords related to generative AI, prompt engineering, and medical writing. The resulting 22 articles form the basis of this review.*
Results: *The evidence demonstrates that* ***prompt engineering****—the methodical design of instructions for an AI—is a critical skill for maximizing the utility and safety of GenAI. Advanced prompting techniques (e.g., Chain-of-Thought, Retrieval-Augmented Generation) are shown to improve the accuracy of data extraction, enhance the readability of lay summaries, and reduce factual "hallucinations." While GenAI significantly accelerates tasks like systematic reviews, the literature reports persistent risks including fabricated references, potential bias, and challenges to traditional authorship. Consequently, there is a unanimous consensus on the non-negotiable need for vigilant human oversight.*
Conclusions: *GenAI offers a powerful adjunct for medical researchers, but its responsible use is not automatic. Effective prompt engineering, coupled with a commitment to ethical guidelines and critical human supervision, is essential to enhance research productivity without compromising scientific integrity.*

Date of Submission: 04-05-2026 Date of Acceptance: 14-05-2026

I. Introduction

The scientific community is in the midst of a profound transformation driven by the advent and rapid integration of generative artificial intelligence (GenAI) [1]. Large language models (LLMs) such as ChatGPT, Gemini, and Claude are unlocking novel avenues to enhance the efficiency and quality of medical research across its entire lifecycle [2, 3]. As detailed in recent reviews, these tools are now being applied to expedite hypothesis generation, streamline systematic literature reviews, draft and refine manuscripts, generate reproducible analysis code, and even support the peer-review process [2, 4, 5]. This technological shift promises to accelerate insight discovery and democratize access to advanced analytical capabilities.

However, the power of these models is unlocked through a critical interface: the user prompt. The literature establishes a clear consensus that the utility and reliability of GenAI are inextricably linked to the skill of **prompt engineering**—the practice of carefully designing instructions to guide the AI toward a specific, accurate, and contextually appropriate outcome [1, 6]. While early interactions with LLMs were often simplistic, a more sophisticated understanding has emerged, recognizing that outputs are highly prompt-dependent and that effective application requires a blend of domain expertise and technical skill [1, 2]. Without deliberate and informed prompting, GenAI tools can produce incorrect, biased, or clinically irrelevant content, posing a significant risk to the integrity of medical research [4].

Given the speed of adoption and the high stakes of medical research, it is essential for the scientific community to have a clear, evidence-based understanding of this new landscape. This narrative review synthesizes the findings from a curated set of 22 recent articles to provide a comprehensive overview of the role of prompt engineering in GenAI-assisted medical research. We will examine the core principles of effective prompting, its measured impact on the quality of research outputs, its application across various research tasks, and the critical challenges and ethical safeguards that must accompany its use. Our objective is to equip researchers, educators, and editors with the knowledge to harness the power of GenAI responsibly, ensuring that this transformative technology augments, rather than undermines, scientific rigor.

II. Method

This narrative review is based on a set of 22 peer-reviewed articles identified through a targeted literature search of the PubMed database. The search was conducted for articles published between 2024 and 2026 using the following specific search string:

("Artificial Intelligence"[Mesh] OR "Natural Language Processing"[Mesh] OR "generative AI"[tiab] OR "large language model*"[tiab] OR "ChatGPT"[tiab]) AND ("prompt engineering"[tiab] OR "prompt design"[tiab]) AND ("Medical Writing"[Mesh] OR "Authorship"[Mesh] OR "scientific writing"[tiab] OR manuscript*[tiab])

The resulting articles form the evidence base for this synthesis.

III. Evidence Synthesis

Core Principles of Effective Prompting

The literature is unequivocal that leveraging GenAI effectively is not an automatic process but a skill rooted in **prompt engineering** [1, 6]. Several reviews identify this as a pivotal competency, noting that outputs are highly prompt-dependent and that a lack of technical expertise in prompting can disadvantage some users [2, 3]. The process involves moving beyond simple commands to providing structured, context-rich instructions. Key techniques highlighted include **Chain-of-Thought (COT)** prompting, which guides the model through a logical sequence, and **Retrieval-Augmented Generation (RAG)**, which grounds the model's output in a specific set of external data or literature, thereby reducing the risk of generating misinformation [4, 7]. To formalize this, researchers have begun developing structured frameworks, such as the **RISEN (Role, Instruction, Steps, End goal, and Narrowing) framework** for creating custom GPTs for data extraction [8], and the **ACUTE (Accuracy, Consistency, Unaltered, Traceability, Ethical) mnemonic** for evaluating the safety of LLM outputs in healthcare [9].

Impact on Key Quality Attributes

The provided literature offers a nuanced view of GenAI's performance, highlighting both remarkable capabilities and significant shortfalls that are modulated by prompt quality. In systematic reviews, LLMs have been shown to achieve high data extraction accuracy (80–94%) but demonstrate only slight-to-moderate agreement in more subjective tasks like risk-of-bias assessment [5]. The risk of **"hallucinations,"** particularly fabricated references, remains a critical quality issue, with rates as high as 47–55% reported in one study [5]. However, the evidence also strongly indicates that performance can be dramatically improved with better prompting and more advanced models. A comparative study on generating lay summaries for prostate cancer manuscripts found that an "extended" prompt design yielded significantly higher readability scores and better alignment with quality thresholds than a "simple" prompt [10]. Similarly, a study on extracting genetic data found that while a baseline model correctly extracted 61% of marker-trait associations, prompt engineering improved this to 91–96% for specific queries [11].

Application of Across Medical Research Tasks

The utility of GenAI, when guided by effective prompting, spans the entire research lifecycle. Several reviews provide a comprehensive overview of applications, including: hypothesis generation, literature synthesis, study design, data analysis, manuscript preparation, and peer review [2, 3, 4]. A major focus of the provided literature is on **systematic reviews**. Studies show that LLMs can streamline this labor-intensive process, with one custom GPT reducing the mean time for data extraction from 36 minutes per study to under 30 seconds of generation plus 13 minutes of human review [8]. However, performance is imperfect. One evaluation found that ChatGPT captured a median of 91% of relevant articles during initial search but that this dropped to 55% during in-depth manuscript screening, reinforcing that LLMs are best used as an adjunct to, not a replacement for, human effort [5, 12, 13]. For **qualitative research**, a step-by-step methodology has been proposed for using a custom GPT to perform thematic analysis based on the Braun and Clarke framework, though it still requires human intervention between steps [14].

Challenges, Risks, and the Imperative for Human Oversight

A powerful and consistent theme across all 22 articles is the explicit acknowledgment of the risks and limitations of GenAI. The primary challenge is the prevalence of **inaccuracies and fabricated content**, which demands that all AI-generated output be treated with skepticism and rigorously verified [2, 5, 13]. Beyond factual errors, the risk of propagating **systemic bias** from training data into research is a key concern [4, 7], as are the critical issues of **patient privacy and data security** when using public, cloud-based AI tools [2, 9]. Furthermore, several papers explore the impact on researchers themselves, warning of **"cognitive offloading"** or the erosion of critical thinking skills that may result from overreliance on AI tools [5, 15]. This connects to the philosophical and practical challenges to **authorship and scientific identity**, with Akgün M. using the "Ship of Theseus"

paradox to question how contribution is defined when human work is progressively replaced by AI output [16]. In response, a clear consensus has formed around the principle of **author accountability**, with human authors remaining fully responsible for the integrity of the work [1, 16]. Ultimately, every paper converges on a single, non-negotiable conclusion: the necessity of **expert human-in-the-loop supervision** to guide, validate, and take ultimate responsibility for the use of GenAI in medical research [1, 5, 8, 12, 13].

IV. Discussion And Future Perspectives

The collective evidence from the provided literature paints a clear and consistent picture: Generative AI is a transformative force in medical research, but it is a "powerful but unstable tool" that requires skillful handling [5]. The synthesis reveals that the central challenge is no longer *if* these tools should be used, but *how* they can be used responsibly to enhance productivity without compromising scientific integrity. The recurring theme across nearly all 22 articles is that **prompt engineering** has moved from a peripheral curiosity to a core competency for the modern researcher [1, 6]. The findings of Rinderknecht et al. [10] and Poretsky et al. [11] provide empirical weight to this, demonstrating a direct, measurable link between prompt sophistication and the quality of AI-generated output.

This reality exposes a critical "skill gap" within the medical community. A cross-sectional study by Maaß et al. reveals that while most medical students are familiar with tools like ChatGPT, they feel unprepared to use them confidently, expressing significant uncertainty about effective prompt engineering and its legal implications [17]. This is compounded by concerns that a reliance on prompting may disadvantage less technical clinicians [2]. This gap highlights an urgent need to integrate formal, structured training on GenAI use, ethics, and prompting into medical education and continuing professional development curricula, a need echoed by multiple authors [15, 17].

Looking forward, the literature points toward several key developments. The focus is shifting from using generic, all-purpose models to developing and validating **domain-specific LLMs and frameworks**. This includes creating practical, step-by-step approaches for safe implementation in healthcare [9], fine-tuning models for specific tasks like semantic understanding [18], and developing custom GPTs for processes like systematic review data extraction [8] and qualitative analysis [14]. Another significant area of research is the potential for GenAI to reform **peer review**. While current models are not yet reliable enough to act as independent reviewers, they show promise in supporting triage tasks and may offer advantages in fairness by reducing human affiliation bias [19, 20]. Ultimately, the future of GenAI in research depends on addressing the limitations identified throughout this review. The unanimous call for rigorous human validation and oversight remains paramount [13]. Future research must continue to quantify the accuracy and bias of these models [2], conduct multi-institutional validations [7], and develop robust ethical frameworks to guide their use [14, 16, 20]. The path forward is not one of full automation, but of an ethically-guided, hybrid model where human expertise directs and validates the powerful capabilities of artificial intelligence.

V. Limitations Of This Review

This review has several limitations that should be acknowledged. First, the literature search was restricted to the PubMed database. While this provides a strong core of biomedical literature, the exclusion of other databases such as Scopus, Embase, and computer science archives like arXiv may have resulted in the omission of other relevant articles. Second, as a narrative review, the synthesis is based on a curated selection of articles and does not follow the exhaustive, systematic methodology of a formal systematic review, which may introduce selection bias. Finally, the field of generative AI is evolving at an unprecedented pace; therefore, this review represents a snapshot in time and may not capture the very latest developments or models that have emerged since the literature search was conducted.

VI. Conclusion

The evidence synthesized in this review confirms that Generative AI is a powerful adjunct for medical researchers, but its responsible and effective use is not automatic. The quality, accuracy, and ethical application of these tools are inextricably linked to the skill of the user in designing effective prompts. Prompt engineering is therefore an essential competency that must be cultivated through formal training and guided by evolving best practices. By embracing a model of critical, human-in-the-loop supervision, the scientific community can harness the productivity gains of GenAI without compromising the integrity and trustworthiness of medical research.

References

[1]. Pu Z, Shi CL, Jeon CO, Et Al. Chatgpt And Generative AI Are Revolutionizing The Scientific Community: A Janus-Faced Conundrum. Imeta. 2024;3(2):E178.
[2]. Kawakita T, Wong MS, Gibson KS, Et Al. Application Of Generative AI To Enhance Obstetrics And Gynecology Research. Am J Perinatol. 2025;42(16):2094-2103.

[3]. Brown JD, Lenchik L, Doja F, Et Al. Leveraging Large Language Models In Radiology Research: A Comprehensive User Guide. Acad Radiol. 2025;32(5):3082-3091.
[4]. Reyes C, Nguyen E, Alexander LF, Et Al. Beyond Human Limits: The Promise And Pitfalls Of Large Language Models In Radiology Research. J Comput Assist Tomogr. 2025;49(4):545-553.
[5]. Gong EJ, Bang CS, Shin YS. Applications Of Large Language Models In Medical Research: From Systematic Reviews To Clinical Studies. Bioengineering (Basel). 2026;13(3):365.
[6]. Venerito V, Lalwani D, Del Vescovo S, Iannone F, Gupta L. Prompt Engineering: The Next Big Skill In Rheumatology Research. Int J Rheum Dis. 2024;27(5):E15157.
[7]. Yang JJ, Hwang SH. Transforming Hematological Research Documentation With Large Language Models: An Approach To Scientific Writing And Data Analysis. Blood Res. 2025;60(1):15.
[8]. Sercombe J, Bryant Z, Wilson J. Evaluating A Customized Version Of Chatgpt For Systematic Review Data Extraction In Health Research: Development And Usability Study. JMIR Form Res. 2025;9:E68666.
[9]. Workum JD, Van De Sande D, Gommers D, Van Genderen ME. Bridging The Gap: A Practical Step-By-Step Approach To Warrant Safe Implementation Of Large Language Models In Healthcare. Front Artif Intell. 2025;8:1504805.
[10]. Rinderknecht E, Schmelzer A, Kravchuk A, Et Al. Leveraging Large Language Models For High-Quality Lay Summaries: Efficacy Of Chatgpt-4 With Custom Prompts In A Consecutive Series Of Prostate Cancer Manuscripts. Curr Oncol. 2025;32(2):102.
[11]. Poretsky E, Blake VC, Andorf CM, Sen TZ. Assessing The Performance Of Generative Artificial Intelligence In Retrieving Information Against Manually Curated Genetic And Genomic Data. Database (Oxford). 2025;2025:Baaf011.
[12]. Yao JJ, Lopez RD, Rizk AA, Aggarwal M, Namdari S. Evaluation Of A Popular Large Language Model In Orthopedic Literature Review: Comparison To Previously Published Reviews. Arch Bone Jt Surg. 2025;13(8):460-469.
[13]. Sollini M, Pini C, Lazar A, Et Al. Human Researchers Are Superior To Large Language Models In Writing A Medical Systematic Review In A Comparative Multitask Assessment. Sci Rep. 2025;16(1):173.
[14]. Cevik AA, Abu-Zidan FM. Utilizing AI-Powered Thematic Analysis: Methodology, Implementation, And Lessons Learned. Cureus. 2025;17(6):E85338.
[15]. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-La-Torre A, Busch F, Ortiz-Prado E. Generative Artificial Intelligence In Medical Education: Enhancing Critical Thinking Or Undermining Cognitive Autonomy? J Med Internet Res. 2025;27:E76340.
[16]. Akgün M. Author Or Prompter? Scientific Writing, Identity, And The Theseus Paradox. Philos Ethics Humanit Med. 2025;20(1):29.
[17]. Maaß L, Grab-Kroll C, Koerner J, Et Al. Artificial Intelligence And Chatgpt In Medical Education: A Cross-Sectional Questionnaire On Students' Competence. J CME. 2024;14(1):2437293.
[18]. Han S, Shi L, Tsui FR. Enhancing Semantical Text Understanding With Fine-Tuned Large Language Models: A Case Study On Quora Question Pair Duplicate Identification. Plos One. 2025;20(1):E0317042.
[19]. Shen SM, Wang Z, Paul K, Li MH, Huang X, Koizumi N. Evaluation Of Large Language Models For Peer Review In Transplantation Research: Algorithm Validation Study. JMIR AI. 2026;5:E84322.
[20]. Grünebaum A, Dudenhausen J, Chervenak FA. The FAIR Framework: Ethical Hybrid Peer Review. J Perinat Med. 2025;53(8):993-999.
[21]. Berry P, Dhanakshirur RR, Khanna S. Utilizing Large Language Models For Gastroenterology Research: A Conceptual Framework. Ther Adv Gastroenterol. 2025;18:17562848251328577.
[22]. Yadav V, Trushna T, Mandal UK, Et Al. AI-Assisted Search Strategy Construction With Step-By-Step Instructions To Execute And Manage Searches Across Major Databases. Med Ref Serv Q. 2026;45(1):1-26.