Commentary

Beyond GPT-4: The Rapidly Evolving Potential of Large Language Models for Clinical Guideline Improvement

Scott D Nelson, PharmD, MS; Adam Wright, PhD

Department of Biomedical Informatics, School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:

Scott D Nelson, PharmD, MS
Department of Biomedical Informatics
School of Medicine, Vanderbilt University Medical Center
3401 West End Ave
Nashville, TN 37203
United States
Phone: 1 6158759347
Email: scott.nelson@vumc.org

Related Article:
Comment on: https://www.jmir.org/2026/1/e81915

Abstract

This commentary reviews the study by Jones et al, which evaluated whether GPT-4 could improve the readability of injectable medication guidelines while preserving important safety information. The study found that GPT-4 produced modest readability gains comparable to manual revision, but also introduced omissions and meaning changes in a minority of sections. These findings highlight both the potential and limitations of early large language models (LLMs) in clinical contexts. However, this study reflects the capabilities of a specific model in a rapidly evolving domain. Since the release of GPT-4, advances in multistep reasoning, model-critique workflows, and structured validation have substantially improved the ability of newer systems to detect omissions, maintain factual fidelity, and support controlled editing. As a result, some documented limitations may stem from the constraints of a single-model, single-pass workflow rather than intrinsic flaws in LLM-assisted guideline revision. This commentary highlights the need for evaluation frameworks that can keep pace with LLM progress and emphasizes that clinical oversight and user-centered testing remain essential. Updated research using contemporary models is needed to determine how emerging architectures can more safely support clarity, consistency, and maintenance of clinical guidelines.

J Med Internet Res 2026;28:e95004; doi: 10.2196/95004

Keywords: artificial intelligence; clinical guidelines; large language model; patient safety; readability; clinical decision support

Introduction

In their recent study, Jones et al [1] evaluated a GPT-4–based pipeline for improving the readability of 20 guidelines from the United Kingdom's National Health Service Injectable Medicines Guide (IMG). The authors found that GPT-4 produced modest but statistically significant readability improvements, and expert pharmacist reviewers rated the revised versions as easier to understand for 26 of 60 (43%) of the ratings. The gains in readability were comparable to those achieved by manual revisions by guideline authors; however, the greatest readability improvements were for two guidelines (aminophylline and voriconazole) using iterative user testing from a previous study.

Notably, using the large language model (LLM) was not without risk. At least one pharmacist reviewer identified omissions in 30 of 153 subsections (20%), additions in 7 subsections (5%), and changes in meaning in 18 subsections (12%). Eight subsections had omissions identified by all 3 reviewers, but no additions or changes in meaning were unanimously flagged. Overall, 65% of all identified issues were flagged by only a single reviewer. The authors concluded that GPT-4 could help augment, rather than replace, manual expert review and user-centered testing to improve guideline readability.

The Challenge of Evaluating a Moving Target

Interpreting these findings requires an appreciation for how rapidly LLM technology has advanced since the release of GPT-4 in March 2023, an extraordinarily long time ago in the current field of artificial intelligence advancements. By the time the study was published, newer models had already introduced major improvements in error checking, multistep reasoning, and structured critique workflows [2,3]. This creates a temporal mismatch where clinical research and guideline development cycles operate on the scale of *years*, while LLM development cycles operate on the scale of *months* and could continue to accelerate [4]. Thus, this study should be viewed as an evaluation of a specific model at a fixed point in time, not a judgment on the overall potential of LLMs in guideline development and review workflows. This is not a criticism of the study itself, as the authors designed a careful, well-controlled evaluation and their findings are valuable precisely because they document specific failure modes that future systems must address. Rather, it is a call to develop more agile evaluation frameworks that can keep pace with technological change, so that evidence generation does not perpetually lag behind the tools available for deployment.

For example, the GPT-4 pipeline relied on a single model for both the editing and quality assurance steps. This architecture creates an inherent tension, since simplification commonly results in removing content, so omissions could be somewhat expected. Newer multiagent systems separate generation from critique, allowing an editor model to propose revisions while a critic model checks for completeness, consistency, and factual accuracy [5]. Structured reasoning frameworks, such as tree-of-thought and self-consistency, enable models to justify edits and cross-check them before finalizing. Skill-based architectures allow explicit function calls to validate medication names, units, values, and section completeness, replacing soft prompts with enforceable programmatic safeguards, while dynamic prompt optimization can iteratively refine instructions to prevent prompt-induced errors. These advances do not eliminate the need for clinician oversight, but they offer more robust mechanisms for preserving informational fidelity while improving readability.

Contextualizing the Use Case

The IMG guidelines represent a relatively favorable use case for LLM revision: procedural, deterministic instructions with clear ground truth. This contrasts with other clinical practice guidelines, such as those for disease management, which must synthesize heterogeneous evidence, navigate gaps, and rely on expert consensus across nonrepresentative study populations, posing far greater challenges. In this example, the study evaluated GPT-4's performance on the IMG, which is a nationally curated, professionally edited resource. Improving readability on an already well-crafted guideline is inherently challenging, yet the model sill produced modest readability improvements. In practice, LLMs may offer even greater value for locally developed clinical guidelines and documents, which are often produced under time pressure with less editorial rigor. Enhancing the clarity and consistency of these documents could improve staff comprehension and ultimately lower the risk of downstream errors. This study tested the LLM against the hardest version of the problem, improving something already well-crafted, while the real-world opportunity may lie in lifting up the documents that need it most.

The Continued Need for User-Centered Testing

User testing remains the guideline improvement technique with the strongest evidence base, and the authors' own data confirm this. Future studies should also include the intended end users. For example, the IMG is primarily used by nurses, who may prioritize quick scanability and visual hierarchy over the pharmacological completeness that pharmacist reviewers would naturally emphasize. The interaction between content accuracy and practical usability can only be fully understood by the people who use these documents at the point of care.

Beyond Readability

Improving readability is only one part of making guidelines more usable. Simplifying text naturally risks omitting information, while providing excessive detail poses its own risks: dense guidelines can cause clinicians to overlook or misinterpret critical information. In the study, 65% of errors were identified by only a single pharmacist, suggesting that many issues were subtle and difficult to detect. The real question is how we can preserve and present essential information in the clearest, most usable form. Newer multimodal models offer approaches beyond text alone. They can generate diagrams, flowcharts, and annotated step-by-step visuals, which may communicate procedural information, such as reconstitution or infusion setup, more effectively than narrative text [6]. These multimodal formats can help reduce cognitive load and help clinicians understand complex instructions more intuitively.

Furthermore, LLMs have broader potential in the guideline ecosystem. They could support translation into other languages, improving access and equity in multilingual care settings [7], though clinical translation would require rigorous verification [8]. LLMs may also enable just-in-time guidance by retrieving and tailoring the relevant portion of a guideline to a clinician's immediate question, which would often be more valuable than improving long documents clinicians may not have time to read. In addition, LLMs could assist with clinical decision support (CDS) maintenance, helping translate updated recommendations into structured CDS logic, or flag conflicts between new evidence and existing rules, reducing alert fatigue and easing the burden of keeping CDS systems current [9]. These applications warrant dedicated study using current-generation models.

Conclusion

Jones et al [1] provide a rigorous, timely evaluation of GPT-4's capabilities and limitations in revising medication guidelines. Their findings identify clear failure modes that future systems must overcome. As LLM architectures continue to advance, updated evaluations are essential to determine how well newer systems address the documented issues and how they can safely support clinicians, guideline authors, and health care organizations. None of these advances eliminate the need for clinician oversight or user-centered testing. The goal is to equip guideline authors and informatics teams with powerful tools for improving how clinical knowledge is communicated and delivered at the point of care. The evidence base must evolve alongside the technology.

Acknowledgments

The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: text generation, proofreading and editing, and summarizing text. The GAI tool used was Microsoft Copilot. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

SDN conceptualized the study and was responsible for writing the original draft. SDN and AW contributed to writing, review, and editing of the manuscript.

Conflicts of Interest

SDN serves on the advisory board for Merative Micromedex and Baxter Healthcare. AW declares no conflicts of interest.

References

1. Jones MD, Torgbi M, Tayyar Madabushi H. Improving the understandability of clinical guidelines: development and evaluation of a GPT-4-based pipeline. J Med Internet Res. Feb 23, 2026;28:e81915. [doi: 10.2196/81915] [Medline: 41730207]
2. Introducing GPT-5. OpenAI. 2025. URL: https://openai.com/index/introducing-gpt-5 [Accessed 2026-04-06]
3. Stanciuc AM. OpenAI's GPT-54 sets new records on professional benchmarks. The Next Web. 2026. URL: https://thenextweb.com/news/openai-gpt-54-launch-computer-use-benchmarks [Accessed 2026-04-06]
4. Aschenbrenner L. Situational Awareness: The Decade Ahead. 2024. URL: https://situational-awareness.ai [Accessed 2026-04-06]
5. Yuan Y, Xie T. Reinforce LLM reasoning through multi-agent reflection. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2506.08379]
6. Benito MD, Diana-Albelda C, García-Martín Á, Bescos J, Viñolo ME, SanMiguel JC. MIRAGE: retrieval and generation of multimodal images and texts for medical education international workshop on applications of medical AI. In: Wu S, Shabestari B, Xing L, editors. Applications of Medical Artificial Intelligence. AMAI 2025. Lecture Notes in Computer Science, Vol 16206. Springer; 2026. [doi: 10.1007/978-3-032-09569-5_11]
7. Pavithra RS. Bridging health literacy gaps in Indian languages: multilingual LLMs for clinical text simplification. In: Zhao W, D'Souza J, Eger S, et al, editors. Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production. Association for Computational Linguistics; 2025. [doi: 10.18653/v1/2025.sciprodllm-1.1]
8. Schlicht IB, Sayin B, Zhao Z, Labonté FM, Barbera C, Viviani M, et al. Disparities in multilingual LLM-based healthcare Q&A. arXiv. Preprint posted online on 2025. [doi: 10.48550/arXiv.2510.17476]
9. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. Jun 20, 2023;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]

Abbreviations

CDS: clinical decision support
IMG: Injectable Medicines Guide

LLM: large language model

Edited by Tiffany Leung; This is a non–peer-reviewed article; submitted 09.Mar.2026; final revised version received 20.Mar.2026; accepted 20.Mar.2026; published 10.Apr.2026

Please cite as:
Nelson SD, Wright A
Beyond GPT-4: The Rapidly Evolving Potential of Large Language Models for Clinical Guideline Improvement
J Med Internet Res 2026;28:e95004
URL: https://www.jmir.org/2026/1/e95004
doi: 10.2196/95004