Utility and Limitations of Large Language Models to Simplify Online Content on Generalized Pustular Psoriasis

SAMER WAHOOD, BA; BENJAMIN GALLO MARIN, MD; OMAR ALANI, ScB; AFUA OFORI-DARKO, BS; DARYA MIREBRAHIMI, BS; KATIE A. O'CONNELL, MD, MS; ARASH MOSTAGHIMI, MD, MPA, MPH; VINOD E. NAMBUDIRI, MD, MBA, EDM

ABSTRACT

Online health information (OHI) in dermatology often exceeds the recommended sixth-grade reading level, hindering patient comprehension. This study aimed to assess the utility of three artificial intelligence large language models (LLMs) – ChatGPT-3.5, ChatGPT-4, and Google Gemini – in enhancing the readability of OHI on generalized pustular psoriasis (GPP) while preserving the reliability and quality of the source material. Texts from the top 20 search results for GPP were reworded by LLMs to a sixth-grade level and evaluated using the enhanced DISCERN instrument and readability indices. Pairwise comparisons of means for each reading scale and DISCERN scores with Tukey's test were also performed. All LLMs significantly reduced readability ($p<0.01$) but scored lower on the DISCERN instrument compared to the original text ($p<0.01$). While LLMs improved readability, they did not preserve the original content's reliability and quality. These findings suggest hesitancy in using LLMs for dermatological patient education.

KEYWORDS: readability; psoriasis; online health information; patient education; large language model

INTRODUCTION

Online health information (OHI) often surpasses the sixth-grade readability level recommendation from the National Institutes of Health.[1] Strategies to improve reading levels of OHI while preserving their meaning have been investigated.[2] Artificial intelligence (AI) large language models (LLMs) are potential tools to aid dermatologists in improving clinical workflow and providing patient education,[3] but their role in simplifying OHI on GPP has not been evaluated. Previously, Malik et al analyzed the original OHI for generalized GPP.[1] Their top 100 search results were entered into commercial LLMs, ChatGPT 3.5, ChatGPT 4.0, and Google Gemini, to evaluate their utility in enhancing readability and preserving the meaning of dermatology-related OHI.[3]

METHODS

Readability and quality were analyzed using the same modified DISCERN instrument and WebFX from Malik et al to compare the original and LLM-produced OHI.[1,3] The modified DISCERN instrument, adapted from Malik et al, is a validated tool for assessing the reliability and quality of online health information.[4] It comprises three core domains: website reliability, treatment information, and disease background. Each website was rated on 5-point Likert scales across items such as clarity of aims, citation transparency, balance, acknowledgment of uncertainty, and comprehensiveness of treatment options – including benefits, risks, and alternatives. Disease-specific items included epidemiology, pathophysiology, symptoms, diagnostic approach, complications, and prognosis. This multifaceted approach enabled a structured comparison of the content accuracy and informational depth between original and LLM-modified educational material.

Infographics, figure legends, videos, and repeated websites were excluded. Each text was entered into the commercial LLMs ChatGPT 3.5, ChatGPT 4.0, and Google Gemini using a new query via the prompt, "Reword this article so an adult with less than a sixth-grade reading level can understand it." Stylistic differences between the original texts and reworded chatbot-generated versions were identified, including the removal of subheadings and omission of medical jargon. All texts were entered on September 5, 2023 to ensure consistency of results and avoid impacts of evolving LLM platforms. Reworded texts were assessed using the tools above and individually compared to their respective originals by three blinded reviewers. Two blinded reviewers then reconciled any differences between the original three reviewers. Each reviewer assessed 100 websites over two LLMs and scored the quality of the information based on modified DISCERN instrument questions. Pairwise comparisons of means for each reading scale with Tukey's honestly significant difference test were performed.

RESULTS

When compared to the original OHI, all three LLMs significantly reduced the average readability grade-level across all three scales ($p<0.01$); however, individual differences between each LLM had no statistically significant difference [**Figure 1**]. Additionally, when compared to the original OHI, all three LLMs scored significantly lower on the DISCERN accountability and treatment scales ($p<0.01$);

however, individual differences between each LLM lacked significance [**Figures 2,3**]. This study uniquely uses a validated scale for the preservation of OHI quality and meaning.

Figure 1. When compared to the original PEMs, all LLMs reduced reading level across all readability indexes with statistical significance (p<0.01).

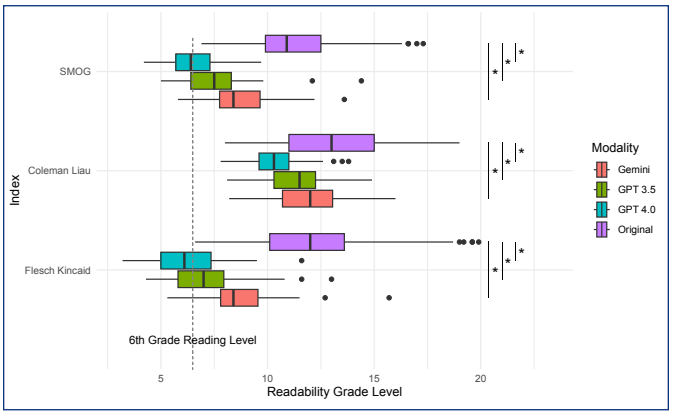


Figure 2. When compared to the original PEMs, all LLMs scored worse on DISCERN accountability scales with statistical significance (p<0.01).

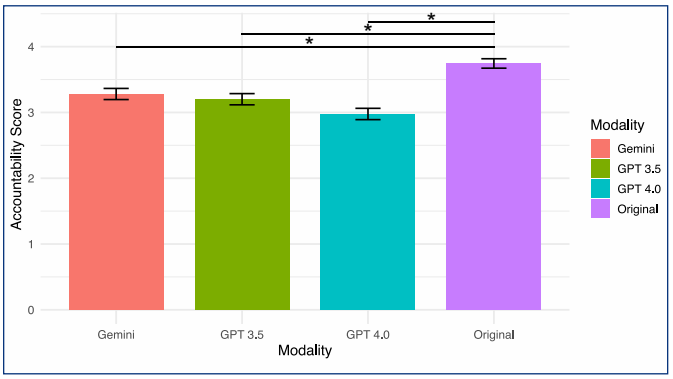
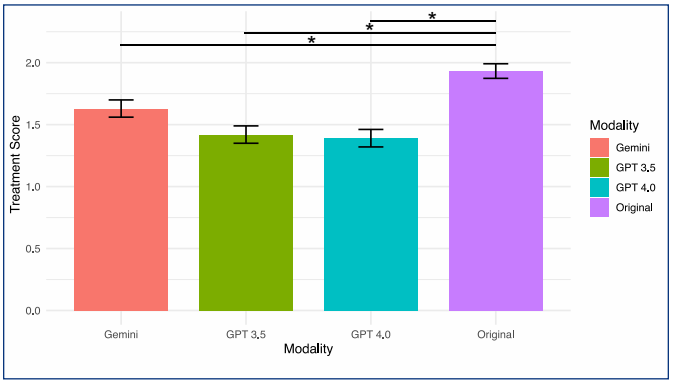


Figure 3. When compared to the original PEMs, all LLMs scored worse on DISCERN treatment scales with statistical significance (p<0.01).



DISCUSSION

While our findings suggest that LLMs may improve readability for dermatologic OHI, they do not preserve the meaning for most sources. This loss of fidelity may stem from the way LLMs generate text. These models rely on predicting the most statistically likely next word rather than verifying factual accuracy, which can lead to "hallucinations," or confident but incorrect or fabricated information.[5] In our study, hallucinations likely occurred during simplification of complex medical content, resulting in omission of important qualifiers, treatment risks, or context. This is a recognized safety concern in clinical settings, where LLMs may offer plausible-sounding but misleading medical advice, misstate diagnostic pathways, or introduce therapeutic inaccuracies.[6] Without mechanisms for real-time source attribution or medical oversight, these limitations highlight the importance of human review before deploying LLM-generated content in patient education.

Although chatbot-generated responses remained above the sixth-grade level, they still represent a step toward improving the accessibility of online health information and addressing educational health inequities. Prompt engineering can be used to optimize responses from LLMs to meet sixth-grade readability.[7] While key concepts were preserved in responses, stylistic differences – such as the removal of subheadings – were identified, suggesting that chatbot responses may not be currently suited to entirely replace human authorship. LLM infographic analysis was not yet available across all LLMs at the time of data collection, hence the research teams' decision to avoid analysis of any infographics. This represents an important limitation in this study that may assist patient comprehension. Nevertheless, our results suggest an emerging role for AI-based interventions in dermatological patient education.

References

1. Malik R, Chen J, Lau C, Sandoval A, Nambudiri VE. Generalized pustular psoriasis: Quality and readability of online health information. Exp Dermatol. 2023;32:1317-1321. doi:10.1111/exd.14775
2. Vallabhaneni A, Eskander PN, Martin K, Eisenstein K, Dyer J. Assessing and optimizing readability of dermatology patient education materials (PEMs). *Pediatr Dermatol*. 2022;39(3):382-384. doi:10.1111/pde.14901
3. Diamond C, Rundle CW, Albrecht JM, Nicholas MW. Chatbot utilization in dermatology: a potential amelioration to burnout in dermatology. *Dermatol Online J*. 2022;28(6). doi:10.5070/D328659734
4. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;53(2):105-111. doi:10.1136/jech.53.2.105
5. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interact J Med Res*. 2025;14:e59823. Published 2025 Jan 28. doi:10.2196/59823
6. Denecke K, May R; LLMHealthGroup, Rivera Romero O. Poten-

tial of Large Language Models in Health Care: Delphi Study. *J Med Internet Res*. 2024;26:e52399. Published 2024 May 13. doi:10.2196/52399

7. White J, Fu Q, Hays S, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint*, Published online February 21, 2023. doi:10.48550/arXiv.2302.11382

Authors

Samer Wahood, BA, The Warren Alpert Medical School of Brown University, Providence, RI.

Benjamin Gallo Marin, MD, Department of Dermatology, Stanford University School of Medicine, Stanford, CA.

Omar Alani, ScB, Icahn School of Medicine at Mount Sinai, New York, NY.

Afua Ofori-Darko, BS, Case Western Reserve University School of Medicine, Cleveland, OH.

Darya Mirebrahimi, BS, Virginia Commonwealth University School of Medicine, Richmond, VA.

Katie A. O'Connell, MD, MS, Department of Dermatology, Vanderbilt University Medical Center, Nashville, TN.

Arash Mostaghimi, MD, MPA, MPH, Department of Dermatology, Brigham and Women's Hospital, Boston, MA.

Vinod E. Nambudiri, MD, MBA, EdM, Department of Dermatology, Brigham and Women's Hospital, Boston, MA.

Acknowledgments

SW and BGM share co-first authorship for this manuscript. Additional collaborators include Waseem Wahood, MD, MS, Kathleen M. Mulligan, MD, and Sarem Rashid, MD. Ethical approval is not applicable for this article.

Disclaimer

The authors report no conflicts of interest relevant to this work. The views expressed herein are those of the authors and do not necessarily reflect the views of their affiliated institutions.

Correspondence

Vinod E. Nambudiri, MD, MBA, EdM
Brigham and Women's Hospital
221 Longwood Ave
Boston, MA, 02115
617-732-4918
Fax 617-582-6060
vnambudiri@bwh.harvard.edu