PRiMER
peer-reviewed reports in medical education research
learner research

Artificial Intelligence-Prompted Explanations of Common Primary Care Diagnoses

Mafaz Kattih | Max Bressler | Logan R. Smith | Anthony Schinelli | Rahul Mhaskar, PhD | Karim Hanna, MD

PRiMER. 2024;8:51.

Published: 9/17/2024 | DOI: 10.22454/PRiMER.2024.916089

Abstract

Background: Artificial intelligence (AI)-generated explanations about medical topics may be clearer and more accessible than traditional evidence-based sources, enhancing patient understanding and autonomy. We evaluated different AI explanations for patients about common diagnoses to aid in patient care.

Methods: We prompted ChatGPT 3.5 , Google Bard, HuggingChat, and Claude 2 separately to generate a short patient education paragraph about seven common diagnoses. We used the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) to evaluate the readability and grade level of the responses. We used the Agency for Healthcare Research and Quality's Patient Education Materials Assessment Tool (PEMAT) grading rubric to evaluate the understandability and actionability of responses.

Results: Claude 2 demonstrated scores of FRE (67.0), FKGL (7.4), and PEMAT, 69% for understandability, and 34% for actionability. ChatGPT scores were FRE (58.5), FKGL (9.3), PEMAT (69% and 31%, respectively). Google Bard scores were FRE (50.1), FKGL (9.9), PEMAT (52% and 23%). HuggingChat scores were FRE (48.7) and FKGL (11.6), PEMAT (57% and 29%).

Conclusion: Claude 2 and ChatGPT demonstrated superior readability and understandability, but practical application and patient outcomes need further exploration. This study is limited by the rapid development of these tools with newer improved models replacing the older ones. Additionally, the accuracy and clarity of AI responses is based on that of the user-generated response. The PEMAT grading rubric is also mainly used for patient information leaflets that include visual aids and may contain subjective evaluations.

Introduction

While use of artificial intelligence (AI) may have wide applications, many are untested, and their capabilities must be examined before implementation into practice.[1-3] While generative AI (GAI) tools have not been formally applied in practice, they are effective in teaching topics like glomerulopathies with an 89% accuracy and are being used to create multiple-choice practice questions by both students and faculty members.[4-6] Furthermore, ChatGPT has been found to be effective at answering first- and second-order medical questions in various topics,[7,8] proving sufficient medical knowledge and passing the multiple-choice United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment, Step 1.[9,10] With these implications in mind, AI may be a tool that can be effectively

utilized in medical education.

Patient education is an essential part of the physician-patient relationship especially when physicians help inform patients of their pathologies. When patients have access to GAI tools, they may gain a better understanding of their diagnoses. Medical students have determined that ChatGPT has clearer and more organized explanations of medical topics than evidence-based sources, although the depth of knowledge of these GAI tools has been limited.[2] It is hypothesized that GAI can generate patient-friendly explanations for common primary care diagnoses, potentially improving patient understanding and engagement. This study aimed to evaluate and compare different AI-generated explanations for common primary care diagnoses that patients may use to better understand these diagnoses.

Methods

We prompted ChatGPT 3.5, Google Bard, HuggingChat, and Claude 2 separately to generate a short patient education paragraph about common diagnoses that included hypertension, hyperlipidemia, type II diabetes mellitus, hypothyroidism, gastrointestinal reflux disease (GERD), atherosclerosis, and vaccination. All four GAI are based on different underlying large language models (LLMs), are free to use, and publicly available. The same prompt was given to each GAI and responses were recorded immediately. Each question was asked in a separate new instance and only the first response was recorded.

Each response was individually assessed using Microsoft Word's Readability Statistic software, which provided a Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). The FRE evaluates text based on how easy it is to read with a score of 60 to 70 being at a level that most adults can easily read and defined as "standard/plain English." The FKGL approximates the grade level of the text. A higher score on the FRE indicates a simpler response, while a lower score on the FKGL indicates a lower grade level. For example, a FKGL of 8 indicates a reading level of around the eighth grade.

The responses were then assessed by the primary investigator using the Agency for Healthcare Research and Quality's (AHRQ) Patient Education Materials Assessment Tool (PEMAT) and its detailed criteria for grading.[11] This tool evaluates the understandability and actionability of patient materials in a systematic way that determines whether patients can understand the material and act based on the material. The form specific for print materials was used. An aggregate score based on the average scores across the seven different diagnoses was determined for each GAI for comparison.

Results

Among the four GAI models, Claude 2 demonstrated the highest average FRE score (67.0) and the lowest FKGL (7.4; Figure 1). Notably, all seven medical diagnosis categories produced by Claude 2 fell within the "standard/plain English" category or easier, representing scores between 60 and 70 (Table 1). ChatGPT exhibited the second highest FRE score (58.5) and FKGL (9.3), with three out of the seven categories meeting the "standard English" category. Google Bard ranked third in readability with an average FRE score of 50.1 and a FKGL of 9.9, however, none of its seven categories fell in the "plain English" range. Lastly, Hugging Chat presented the least readable text (48.7 FRE and 11.6 FKGL).

Table 2 shows the assessment of understandability and actionability using the PEMAT. Claude 2 again demonstrated the highest combined scores, averaging 69% for understandability and 34% for actionability (Figure 2). ChatGPT followed with the next highest understandability and actionability scores of 69% and 31% respectively, while Hugging Chat scored 57% for understandability and 29% for actionability. Google Bard presented the lowest scores, registering 52% for understandability and 23% for actionability.

Conclusions

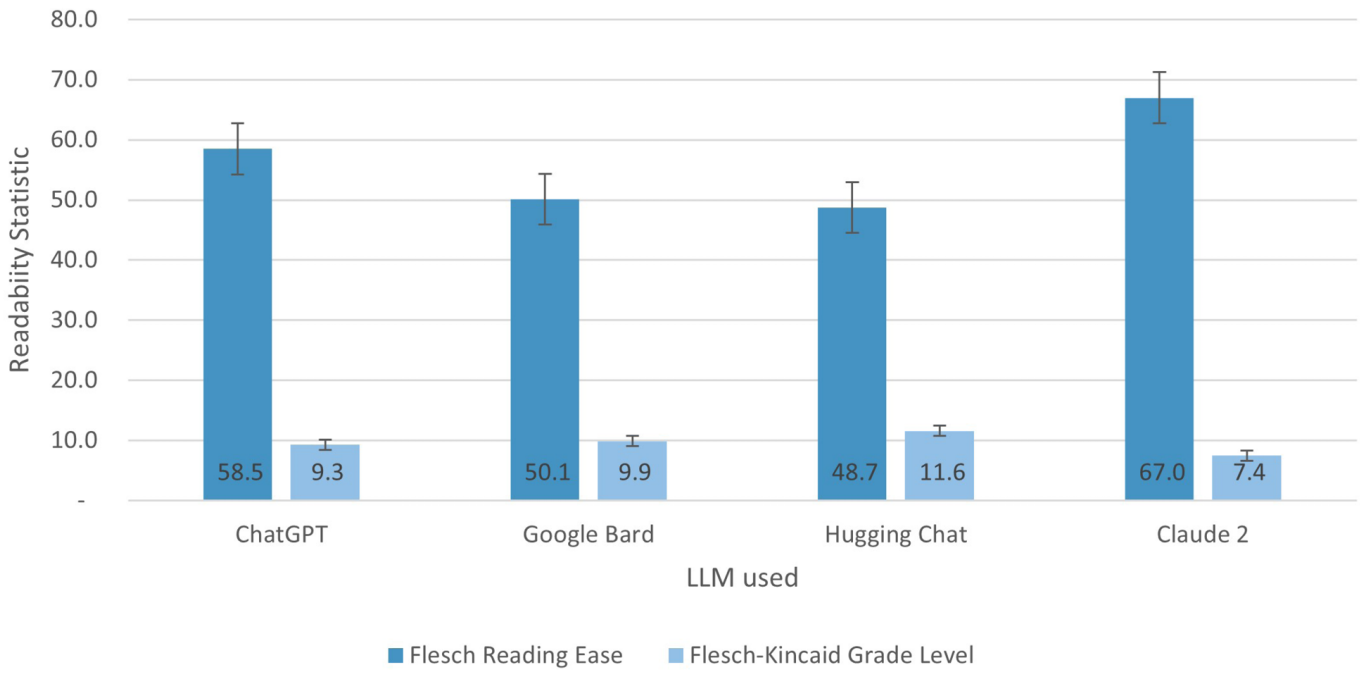
GAI services can provide an opportune way for patients to enhance their autonomy with a strong understanding of their disease. Previous studies found that ChatGPT creates significantly easier questions than Bard or Bing AI,[12] however regarding the readability and accessibility of patient explanations, the differences between the GAI may not be significant. While Claude 2 and ChatGPT demonstrated superior readability and understandability, practical application and patient outcomes need further exploration. Furthermore, three of the four GAI were able to generate explanations below the level of a tenth grader. Overall, this study demonstrates that the availability of GAI models contributes to their utility for educating patients about their diagnoses, both by providers as well as by the patients themselves. These AI tools can be integrated into patient care by providing explanations during consultations or as part of after-visit notes, thus reinforcing patient understanding and adherence to treatment plans. When deciding which GAI to use, it is important to consider both the readability of the text that will be provided as well as the reading level of the target audience.

By placing more emphasis on self-directed and resource-driven learning that utilizes AI for efficient synthesis of knowledge from different sources, both patient autonomy and quality of care could increase.[13] While the use of AI in medicine carries a host of ethical and logistical challenges, it can also enhance a patient's ability to learn about their own disease.[1] There is an ongoing conversation about the application of these tools into medical education with one possible route including their use to augment current education methods.[14] Some postulate that the threats of GAI– including overreliance, accuracy, bias, and the problem of hallucinations (GAI generating false information to answer a prompt) must be heavily considered and the use of GAI must be heavily supervised.[15-18] To mitigate these risks, it may be useful refer to published AI competencies that can help evaluate the tools in a controlled manner to prevent misuse in education.[3,16] While these GAI can provide patients with an adequate learning tool, they may impact the physician-patient relationship. Two possible ways to ensure positive AI use are to maintain GAI in an assistant role and adapt medical education to include AI competence.[19,20]

This study had several limitations, perhaps the biggest being the rapid pace of development of AI. As new models continue to be trained and released for public use, the older models from which this investigation is based on may no longer be in use. In addition, the accuracy and clarity of AI responses is based on the prompt's own clarity and wording.[2] Our study prompts were controlled by asking all four different GAI the same prompt with new instances all on the same day. Another limitation lies in the PEMAT grading rubric. The AI responses were assessed by the primary researcher and while attempts were made to follow the grading criteria as closely as possible, there may be some subjective bias in how certain aspects of the responses were rated. Furthermore, the reliability of the explanations was not assessed other than a first look to confirm the explanations did not contain any false information. Further study is needed to evaluate the real-world impact of AI-generated patient education materials on patient knowledge retention, behavior change, and overall health.

Tables and Figures

Figure 1. Average Flesch Reading Ease and Flesch-Kincaid Grade Level of Different LLMs Based on Microsoft Word Readability Statistic



Abbreviation: LLM, large language model

Table 1. Flesch Reading Ease and Flesch-Kincaid Grade Level for Different LLM Responses Based on Microsoft Readability Statistics

	ChatGPT		Google Bard		Hugging Chat		Claude 2	
	Flesch Reading Ease	Flesch-Kincaid Grade Level	Flesch Reading Ease	Flesch-Kincaid Grade Level	Flesch Reading Ease	Flesch-Kincaid Grade Level	Flesch Reading Ease	Flesch-Kincaid Grade Level
Hypertension	70.6	8.5	54.8	9.3	44.3	12.3	65.5	8.3
Hyperlipidemia	59.6	9.6	41.0	11.1	53.9	10.8	62.5	8.6
Type 2 diabetes	61.4	9.4	52.4	9.8	51.7	11.4	70.6	6.6
Hypothyroidism	44.0	11.5	50.1	9.8	59.4	8.9	69.1	6.4
GERD	54.1	9.3	46.7	10.1	43.8	13.3	66.4	6.6
Atherosclerosis	59.2	8.3	58.3	8.7	50.4	10.8	70.3	7.2
Vaccination	60.5	8.5	47.6	10.4	37.4	13.6	64.5	8.4
Average	58.5	9.3	50.1	9.9	48.7	11.6	67.0	7.4
Standard deviation	8.06	1.10	5.70	0.77	7.35	1.63	3.09	0.96

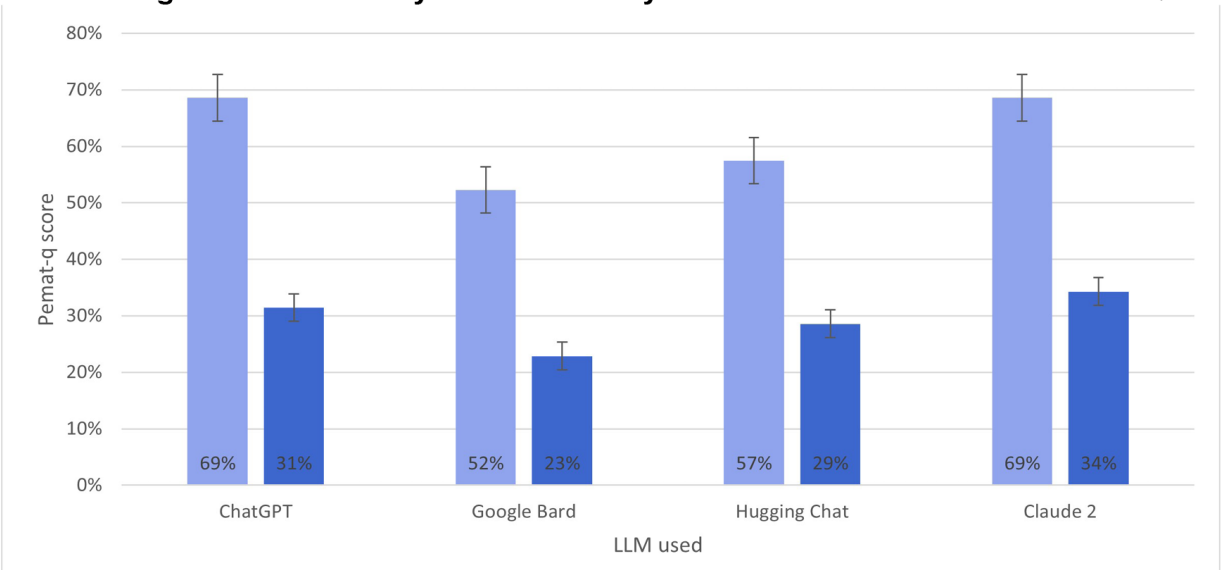
Abbreviations: LLM, large language model; GERD, gastrointestinal reflux disease

Table 2. Understandability and Actionability of Different LLM Responses Based on PEMAT-Q Rubric

	ChatGPT		Google Bard		Hugging Chat		Claude 2	
	Understandability	Actionability	Understandability	Actionability	Understandability	Actionability	Understandability	Actionability
Hypertension	67%	40%	33%	40%	67%	40%	67%	20%
Hyperlipidemia	67%	40%	56%	20%	67%	60%	67%	60%
Type 2 diabetes	67%	40%	44%	20%	56%	20%	67%	20%
Hypothyroidism	67%	40%	44%	20%	56%	20%	78%	60%
GERD	67%	20%	44%	20%	56%	20%	67%	40%
Atherosclerosis	67%	20%	67%	20%	56%	20%	56%	20%
Vaccination	78%	20%	78%	20%	44%	20%	78%	20%
Average	69%	31%	52%	23%	57%	29%	69%	34%
Standard deviation	4%	11%	16%	8%	8%	16%	8%	19%

Abbreviations: PEMAT-Q, Patient Education Materials Assessment Tool Question; GERD, gastrointestinal reflux disease.

Figure 2. Average Understandability and Actionability of Different LLMs Based on PEMAT-Q Rubric



Abbreviations: LLM, large language model; PEMAT-Q, Patient Education Materials Assessment Tool question

Acknowledgments

The authors declare no conflicts of interest related to this study. AI tools were not used in writing this article.

Corresponding Author

Karim Hanna, MD

Morsani College of Medicine, University of South Florida, Tampa, FL;Department of Family Medicine, Morsani

College of Medicine, University of South Florida, Tampa, FL
khanna@usf.edu

Author Affiliations

Mafaz Kattih - Morsani College of Medicine, University of South Florida, Tampa, FL
Max Bressler - Morsani College of Medicine, University of South Florida, Tampa, FL
Logan R. Smith - Morsani College of Medicine, University of South Florida, Tampa, FL
Anthony Schinelli - Morsani College of Medicine, University of South Florida, Tampa, FL
Rahul Mhaskar, PhD - Morsani College of Medicine, University of South Florida, Tampa, FL
Karim Hanna, MD - Morsani College of Medicine, University of South Florida, Tampa, FL | Department of Family Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL

References

1. van der Niet AG, Bleakley A. Where medical education meets artificial intelligence: 'Does technology care?'. *Med Educ*. 2021;55(1):30-36. doi:10.1111/medu.14131
2. Breeding T, Martinez B, Patel H, et al. The utilization of ChatGPT in reshaping future medical education and learning perspectives: a curse or a blessing? *Am Surg*. 2023;•••:31348231180950.
3. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. *Ann Fam Med*. 2022;20(6):559-563. doi:10.1370/afm.2887
4. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. *JMIR Med Educ*. 2022;8(2):e35587. doi:10.2196/35587
5. Hamdy H, Sreedharan J, Rotgans JI, et al. Virtual Clinical Encounter Examination (VICEE): A novel approach for assessing medical students' non-psychomotor clinical competency. *Med Teach*. 2021;43(10):1203-1209. doi:10.1080/0142159X.2021.1935828
6. Cross J, Robinson R, Devaraju S, et al. Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. *Cureus*. 2023;15(7):e41399. doi:10.7759/cureus.41399
7. Das D, Kumar N, Longjam LA, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*. 2023;15(3):e36034. doi:10.7759/cureus.36034
8. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. *Cureus*. 2023;15(4):e37023. doi:10.7759/cureus.37023
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
10. Gilson A, Safranek CW, Huang T, et al. How does chatgpt perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. doi:10.2196/45312
11. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns*. 2014;96(3):395-403. doi:10.1016/j.pec.2014.05.027
12. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*. 2023;15(6):e40977. doi:10.7759/cureus.40977
13. Cutrer WB, Spickard WA, 3rd, Triola MM, Allen BL, Spell N, 3rd, Herrine SK, et al. Exploiting the power of

information in medical education. *Medical Teacher*. 2021;43(sup2):S17-S24. doi:10.1080/0142159X.2021.1925234
14. Hanna K. Exploring the applications of ChatGPT in family medicine education: five innovative ways for faculty integration. *PRiMER Peer-Rev Rep Med Educ Res*. 2023;7:26. doi:10.22454/PRiMER.2023.985351
15. Horton JA, Ally I. Response to "exploring the applications of chatgpt in family medicine medical education". *PRiMER Peer-Rev Rep Med Educ Res*. 2023;7:28. doi:10.22454/PRiMER.2023.940827
16. Liaw W, Chavez S, Pham C, Tehami S, Govender R. The hazards of using chatgpt: a call to action for medical education researchers. *PRiMER Peer-Rev Rep Med Educ Res*. 2023;7:27. doi:10.22454/PRiMER.2023.295710
17. Kleebayoon A, Wiwanitkit V. The hazards of using chatgpt: additional comments. *PRiMER Peer-Rev Rep Med Educ Res*. 2024;8:12. doi:10.22454/PRiMER.2024.203721
18. Armitage RC. ChatGPT: the threats to medical education. Postgraduate Medical Journal. 2023;06:06. doi:10.1093/postmj/qgad046
19. Watts E, Patel H, Kostov A, Kim J, Elkbuli A. The role of compassionate care in medicine: toward improving patients' quality of care and satisfaction. *J Surg Res*. 2023;289:1-7. doi:10.1016/j.jss.2023.03.024
20. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak*. 2023;23(1):73. doi:10.1186/s12911-023-02162-y