Evaluating the Accuracy and Quality of ChatGPT-4o Respones to Patient questions on Reddit

Mihailo Svetozarević†
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
mihailo.svetozarevic@gmail.com

Isidora Svetozarevic
Center for Radiology
University Clinical Center Niš
Niš, Serbia
isidora_jankovic@yahoo.com

Sonja Jankovic
Center for Radiology
University Clinical Center Niš
Niš, Serbia
sonjasgirl@gmail.com

Stevo Lukić
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
srlukic@gmail.com

Abstract

The rapid integration of large language models (LLMs) into healthcare communication has raised questions about their accuracy, safety, and usefulness for patients seeking medical advice online. This study evaluated the performance of ChatGPT-4o in responding to epilepsy-related patient questions posted on the r/AskDocs subreddit. A total of 110 questions were selected based on the keywords epilepsy, seizure, and seizure disorder, filtered by the "physician responded" flair. Responses generated by ChatGPT-4o were independently assessed by four physicians across multiple domains including accuracy, comprehensiveness, clarity, relevance, and empathy as well as binary assessments of bias, factuality, fabrication, falsification, plagiarism, harm, reasoning, and currency. Results showed that most of the responses were rated as good or very good, with particularly high scores for accuracy, clarity, relevance, and comprehensiveness, while empathy was consistently lower. These findings suggest that ChatGPT-4o may serve as a useful complementary tool for patient education and engagement in epilepsy, though it cannot replace professional medical consultation. Future research should further investigate its role in clinical practice and strategies for improving empathetic communication in AI generated responses.

Keywords

ChatGPT-4o, epilepsy, seizure disorder, artificial intelligence, patient communication, evaluation, accuracy, empathy, large language models

1. Introduction

In medicine, large language models (LLMs) are increasingly applied to diverse tasks, including information extraction from electronic health records, scientific writing support, patient care documentation, and even clinical guideline development. Importantly, the use of LLMs is not limited to healthcare professionals. Patients themselves are increasingly experimenting with these tools, as new models and updated versions create the impression of rapidly expanding capabilities from one year to the next. This steady rise in LLM use coincides with an already well-established pattern: health information is often sought online before consulting a physician. In the United States, survey data show that about six in ten adults aged 18 to 29 report being online almost constantly, with somewhat smaller but still substantial proportions in older groups. Such an environment directly encourages digital health information-seeking behavior and frequent encounters with LLM-based tools. [1,2]

The COVID-19 pandemic further accelerated the adoption of virtual health care and normalized the use of public online forums where patients seek advice sometimes from reliable professionals, but often from peers or unverified sources. Reddit, along with similar platforms, has become a representative setting for "real-world" patient - physician interactions in an asynchronous, text-based format. The potential advantages of LLMs in this context are considerable. They can rapidly synthesize information, explain disease mechanisms in accessible language, highlight red-flag symptoms, and point to relevant resources, all while being available around the clock. They are also generally intuitive to use, even for individuals with limited health literacy. Furthermore, recent evaluations suggest that LLM-generated responses may convey greater empathy and clarity than physician-written answers in some online settings, potentially improving comprehension and adherence. Yet, the risks remain substantial. LLMs are prone to generating hallucinations plausible but incorrect statements

while omitting key information or inferring unstated details. In a high-risk domain such as medicine, these limitations render unsupervised use unsafe. The most recent literature emphasizes that hallucinations and omissions are intrinsic to current LLM architectures, and that without rigorous safeguards - such as benchmarking, oversight, and validation - clinical deployment should not proceed unchecked. Beyond technical concerns, the rapid spread of LLM use also raises new ethical and societal challenges. Healthcare is guided by strict ethical norms, professional duties, and societal responsibilities, and recent case reports highlight instances where LLM outputs, including those from ChatGPT, have contributed to harmful and potentially life-threatening outcomes. [3]

In this review, we focus on a specific clinical domain - epilepsy and other seizure disorders where the need for reliable information is particularly acute. Epilepsy is a chronic, often lifelong condition with a heterogeneous clinical presentation, typically beginning in childhood or young adulthood. Patients with epilepsy frequently have questions about treatment options, drug interactions, lifestyle considerations, and safety precautions. Studies have shown that a significant proportion of individuals with epilepsy actively search for information online, both on general and disease-specific topics. Analyses of search patterns (for example, on Wikipedia) have revealed strong public interest and episodic peaks in epilepsy-related queries. More recent research indicates that people with epilepsy engage in online health information seeking at higher rates than many other patient groups, underscoring the importance of understanding how LLM responses might influence their perceptions and behaviors. However there are both potential benefits and inherent limitations of LLMs in epilepsy care as shown by recent review articles. [4,5,6,7,8,9]

Despite the growing body of literature on LLMs in medicine, they remain insufficiently reliable for routine, uncontrolled use. A notable gap exists: few studies evaluate LLMs from the patient's perspective, particularly using real-world data drawn from public forums. Our study is designed to address this gap. Specifically, we assess whether responses generated by OpenAI's ChatGPT-4 meet the needs of people with epilepsy who ask questions on r/AskDocs. Physicians serve as expert evaluators not to arbitrate "on behalf of patients," but to operationalize criteria of quality, utility, accuracy, and safety in line with real user needs. We argue that this design places the patient - LLM relationship at the center of the analysis, while leveraging medical expertise to standardize evaluation metrics and identify areas where safeguards or clinical verification remain necessary. In this framework, Reddit provides a natural, heterogeneous, and timely source of patient queries, enabling an evaluation of LLM responses under conditions that approximate the realities of everyday patient information-seeking. . [3,7,10,11]

2. Material and Method

In the intial phase of the study we collected a total of 110 patient questions from the subreddit r/AskDocs, one of the more active medical communities on reddit with over half a milion active participants. Questions were identified using a filtered search using keywords „epilepsy", „seizure" and „seizure disorder". To ensure quality only posts submitted within the past 12 months and those that received at least one verified physican response (marked with the flair „physician responded") were included. Out of the selected 110 questions, 4 were excluded due to being duplicates or irrelevant to the subject matter.

For each selected question a response was generetad using ChatGPT 4.0. These responses were then independently evaluated by four certified physicians – one neurologist, one radiologist, one neurology resident and one radiology resident. The raters were blinded to each other's assesments and did not consult each other during the evaulation process. Interrater agreements were reached using Fleiss Kappa with minimal discrepanies observed among evaluators.

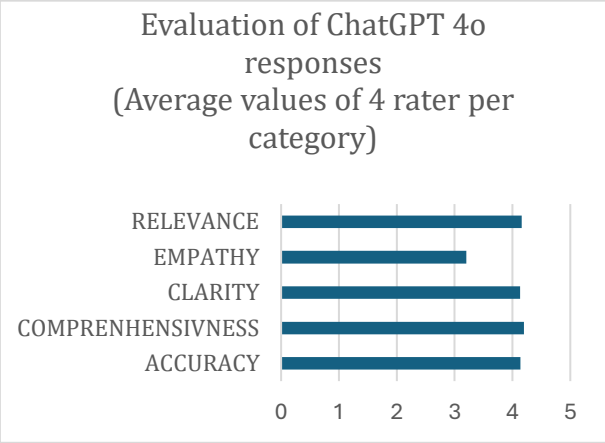
Evaluations were made using predefined dimension with a modified Likert scale (1-5). The dimensions assessed were Accuracy, Comprehensiveness, Clarity, Empathy, Relevance. Addidional dimensions were assessed using categorical ratings (Yes/No responses). These dimensions were Reasoning, Currency, Bias, Harm, Factuality, Fabrication, Falsification and Plagiarism.

3. Results

Overall the raters found that ChatGPT 4.0 respones were very positive with approximately 80% of answers classified as „good" or „very good" across all dimensions on the Likert Scale. Most answers were considered factually correct, we found no responses to be incorrect. Most answeres were very thourough and easily understandable with language that the raters believe cover all educational specters. We found no instanes of outdated recommendations and all responses were deemed to be concise, without unnecessary and overbearing details.

Regarding categorical measures we did not find any cases of bias, harm, fabrication, falsification. All answers gave information that could be easily varified against standard medical sources. The lowest scoring dimension was empathy as we found most answers to be on average good or decent with no responses being explicitly poor.

All together, these results suggest that ChatGPT 4.0 is capable of generating accurate, clear and relevant responses to patient questions about epilepsy with the primary limitation being in the domain of empathetic responses.



4. Discussion

In this study, we examined the usability of responses generated by ChatGPT 4.0 in comparison to neurologists' answers to patient questions about epilepsy on Reddit, specifically the subreddit r/AskDocs. This community is one of the largest and most active health forums online, with over half a million members and hundreds of new patient questions submitted daily. A particular strength of this platform lies in its anonymity: users can ask sensitive medical questions more openly than they might in a clinical encounter, which results in a broader and more candid spectrum of concerns. Additionally, r/AskDocs is actively moderated and follows strict rules medical advice is permitted only from verified physicians (marked by a special flair), while other users are restricted to sharing personal experiences. This structure ensures a basic level of quality control and provides a reliable basis for comparing physician responses with those of ChatGPT. We believe this makes r/AskDocs a relevant and valid environment for evaluating the potential of large language models (LLMs) in a medical setting.

Our findings complement recent research done by Fennig and colleages [12], in which LLM models were used to analyze tens of thousands of Reddit posts to identify topics and concerns that epilepsy patients often do not bring up in clinical settings. That work found significant patterns such as stigma, emotional distress, substance use, and seizure description high-engagement topics that are outside of standard outpatient conversations and often not given adequate space in the clinical conversation. This confirms that LLM models are not only for providing answers, but also for a deeper understanding of patient needs, which further justifies the use of r/AskDocs as a source of realistic and relevant questions for our study.

Our findings indicate that ChatGPT 4.0 generally provides accurate, relevant, and comprehensive answers. Importantly, no response was deemed explicitly incorrect, underscoring the potential of such tools to deliver reliable medical information for patients with epilepsy. However, the model consistently showed weaker performance in conveying empathy compared to physicians. This limitation has been noted in previous studies, which emphasize that while LLMs can reproduce medical content accurately, they struggle to replicate the human aspects of communication such as reassurance, compassion, and emotional support. [1,6,8]

The overall impression of the neurologists was that the ChatGPT 4.0 responses were mostly "acceptable" or "good", while a smaller number were rated as "very good". Nevertheless, doctors generally gave somewhat better answers, but the difference was not large. This finding is consistent with the results of a study by Ayers and colleagues., who also found that chatbot responses can be of similar or even better quality in certain dimensions, but with limitations in empathy. [1]

It is important to point out that our results should be seen in the context of the increasing number of patients using the Internet for epilepsy information and potentially changing therapy based on information obtained online. Previous studies have shown that patients with epilepsy frequently search the Internet to learn more about their disease [3,4], while more recent studies indicate a high rate of use of digital sources of health information in this population. [5] Precisely because of this, the ability of large language models to generate correct and comprehensible answers is of particular importance.

Even though our findings are encouraging, it is necessary to emphasize the potential risks. The literature on LLMs in medicine warns of the phenomenon of "hallucinations", i.e. giving confident but incorrect answers. [6,7] Although in our series no answer was explicitly wrong, such cases were not excluded in a larger sample, especially in more complex clinical scenarios. In addition, a critical review of LLMs in epileptology indicates that current tools may be useful for patient and physician education, but are not ready for routine, uncontrolled clinical application. [8]

Finally, it should be emphasized that the focus of our study was the attitude of patients towards the responses of LLMs, while doctors had the role of mediators in quality assessment. This kind of perspective can be significant for future research, as it opens up space for a better understanding of how patients value and perceive such tools compared to traditional medical sources

5. Conclusion

This study demonstrates that ChatGPT 4.0 provides responses to patient questions about epilepsy that are largely accurate, relevant, clear, and comprehensive. However, the limitations observed - especially regarding emotional support and nuanced communication highlight that ChatGPT cannot replace professional medical consultation. Instead, its role should be considered complementary, supporting patient education and engagement, while final interpretation and guidance remain within the responsibility of qualified healthcare professionals.

Acknowledgements

Views and opinions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor any other authority can be held responsible for them. All authors contributed equally in the final version of this paper.

References

1. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. doi: 10.1001/jamainternmed.2023.1838
2. Pew Research Center. Mobile technology and home broadband 2021. Available from: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/
3. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). 2025;5:44. doi: 10.1038/s43856-025-01021-3
4. Liu J, Dong X, Mao Y, et al. Internet usage for health information by patients with epilepsy. Epilepsy Behav. 2013;29(1):110–3. doi: 10.1016/j.seizure.2013.06.007
5. Brigo F, Erro R, Marangi A, et al. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy Behav. 2015;45:128–33. doi: 10.1016/j.yebeh.2013.11.020
6. Bingöl N, Mutluay FK, Erbaş O. Determining the health-seeking behaviors of people with epilepsy. Epilepsy Behav. 2024;152:109331. doi: 10.1016/j.yebeh.2024.110063
7. Bélisle-Pipon JC, et al. Why we need to be careful with large language models in medicine and healthcare. AI & Ethics. 2024. doi: 10.3389/fmed.2024.1495582
8. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of large language models for medical text summarization. npj Digit Med. 2025;8(1):19. doi: 10.1038/s41746-025-01670-7
9. García-Azorín D, Bhatia R, et al. Potential merits and flaws of large language models in epilepsy care: A critical review. Epilepsia. 2024;65(4):873–886. doi: 10.1111/epi.17907
10. The Verge. Google's healthcare AI made up a body part. The Verge. 2025 May 7. Available from: https://www.theverge.com/2025/05/07/google-healthcae-ai-made-up-body-part
11. Auvin S, Nabbout R, et al. Quality of health information about epilepsy on the Internet. Arch Pediatr. 2013;20(6):603–7. doi: 10.1016/j.neurol.2012.08.008
12. Fennig U, Yom-Tov E, Savitzky L, Nissan J, Altman K, Loebenstein R, Boxer M, Weinberg N, Gofrit S G, Maggio N. Bridging the conversational gap in epilepsy: Using large language models to reveal insights into patient behavior and concerns from online discussions. Epilepsia. 2025;66(3):686–699. doi: 10.1111/epi.18226