Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: A Comparison of Performances on the Responses to Common Queries.

Linger Li, Pengfei Li, Kun Wang, Liang Zhang, Hongqin Zhao, Hongwei Ji

Submitted to: Journal of Medical Internet Research
on: December 29, 2023

Table of Contents

Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: A Comparison of Performances on the Responses to Common Queries.

Linger Li[1] MM; Pengfei Li[2] MD; Kun Wang[1] MM; Liang Zhang[1] MD; Hongqin Zhao[1*] MD, PhD; Hongwei Ji[3, 4*] MD

[1]Department of Neurology The Affiliated Hospital of Qingdao University Qingdao CN
[2]Department of Internal Medicine The Affiliated Hospital of Qingdao University Qingdao CN
[3]Tsinghua Medicine Tsinghua University Beijing CN
[4]Department of Internal Medicine Beijing Tsinghua Changgung Hospital Beijing CN
[*]these authors contributed equally

Corresponding Author:
Hongqin Zhao MD, PhD
Department of Neurology
The Affiliated Hospital of Qingdao University
No.59 Haier Road
Qingdao
CN

Abstract

Migraine, a frequent and highly disabling disorder, necessitates enhanced education of individuals with migraine to mitigate this global burden. The rapidly evolving field of large language models (LLMs) presents a promising avenue for assisting in migraine patient education. This study aims to assess the potential of LLMs in this context by evaluating the accuracy of responses from five leading LLMs, including OpenAI's ChatGPT 3.5 and 4.0, Google Bard, Meta Llama2, and Anthropic Claude2, in addressing 30 commonly asked migraine-related queries. We found that LLMs demonstrated varied levels of accuracy. ChatGPT-4.0 provided 96.7% appropriate responses, while other chatbots provided 83.3% to 90% appropriate responses. This study underscores the potential of LLMs, notably ChatGPT-4.0, demonstrated to accurately address common migraine-related queries. Such findings could advance AI-assisted education for individuals with migraine, providing insights for a holistic approach to migraine management.

(JMIR Preprints 29/12/2023:55927)
DOI: https://doi.org/10.2196/preprints.55927

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
✓ **Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a h**

Original Manuscript

Benchmarking State-of-the-Art Large Language Models for Migraine Patient Education: A Comparison of Performances on the Responses to Common Queries.

Linger Li[1], Pengfei Li[2], Kun Wang[1], Liang Zhang[1], Hongqin Zhao[1], Hongwei Ji[3,4]

1. Department of Neurology, The Affiliated Hospital of Qingdao University, Qingdao 266035, China.
2. Department of Internal Medicine, The Affiliated Hospital of Qingdao University, Qingdao 266035, China.
3. Tsinghua Medicine, Tsinghua University, Beijing 100084, China.
4. Department of Internal Medicine, Beijing Tsinghua Changgung Hospital, Beijing 102218, China.

Correspondence:

Prof. Hongqin Zhao, Department of Neurology, The Affiliated Hospital of Qingdao University, No.59 Haier Road, Qingdao 266035, Shandong, China. Phone: 13864873935. Email: zhaohongq@qdu.edu.cn.

Paper type: Research Letter

Keywords: migraine; large language models; patient education; ChatGPT; Google Bard.

Introduction

Migraine is a highly debilitating primary headache disorder. Long-term patient education for migraine is essential to help patients, identify triggers, use medications appropriately, and adopt lifestyle changes that can reduce the frequency and severity of attacks[1-3]. Large language models (LLMs), which are complex artificial intelligence systems trained on extensive datasets to produce human-like text responses, have emerged as a promising tool for patient education. However, LLMs could lead to inaccurate responses, known as 'hallucinations'[4]. Therefore, rigorous evaluations within specific medical domains are essential. Nevertheless, there is a lack of benchmarking for popular online LLMs in migraine patient education. This study conducted a comparative analysis of responses from five leading LLMs, namely OpenAI's ChatGPT 3.5 (December, 2022) and 4.0 (March, 2023), Google Bard (February, 2023), Meta Llama2 (July, 2023), and Anthropic Claude2 (July, 2023).

Methods

This study was conducted from October 1 to October 28, 2023. Neurologists, utilizing guidelines and their clinical expertise, meticulously crafted 30 migraine-related queries. These queries spanned a wide array of topics, including evaluation and definition, testing and diagnosis, treatment, follow-up and prognosis, as well as special population considerations (*Table 1*)[5-7]. Each query was individually presented to five leading LLMs through an independent conversation interface via a web-based interface to avert chained prompting. Responses from all LLM-Chatbots were converted into plain text, obscuring specific chatbot features to ensure blinding. The generated answers were then randomly ordered in each set of queries and assessed by blinded reviewers, with a 24-hour interval between each assessment to mitigate memory bias (*Figure 1*).

Table 1. Performance of Large Language Models in Addressing Patient Queries.

	GPT-3.5	*GPT-4*	*Bard*	*Llama 2*	*Claude2*	*P value*

Appropriate response,[a] n (%)	27(90.0)	29(96.7)	25(83.3)	27(90.0)	26(86.7)	0.481
Poor response,[b] n (%)	1(3.3)	1(3.3)	2(6.7)	1(3.3)	1(3.3)	0.961
1. Evaluation and definition						
Why am I experiencing migraines?	9	9	9	9	9	-
Is migraine a common disease?	9	9	9	9	9	-
Is there an association between stress and migraine?	9	9	9	9	9	-
Is there an association between exercise and migraine?	9	9	9	9	9	-
Is there an association between patent foramen ovale and migraine?	9	9	9	9	9	-
I have migraine, does this have anything to do with my family's health history?	9	9	9	9	9	-
What is the best diet for migraine patients?	9	9	9	9	9	-
Why is my vision blurred before the migraine attack?	8	9	9	9	8	-
Why do I get frequent migraine attacks?	9	9	9	9	9	-
Why do I feel dizziness when I have a migraine attack?	9	9	7	8	9	-
Why do I have migraine attacks during menstruation?	9	9	9	9	9	-
2. Test and Diagnosis						
How is migraine diagnosed?	9	9	9	9	9	-
Is my migraine serious? how to evaluate it?	9	9	8	9	9	-
I have a migraine, is it a tumor in my brain?	8	9	9	8	8	-
How do imaging tests like MRI, CT scans, or ultrasounds contribute to diagnosing migraines?	9	9	7	9	9	-
3. Treatment						
How is migraine treated?	9	9	9	9	9	-
Do migraines require long-term medication therapy?	9	9	9	9	9	-
I have a migraine, what will happen if I don't treat it?	9	9	9	9	9	-
What do migraine patients need to pay attention to when taking medicine?	9	9	9	9	9	-
What lifestyle changes can I make to better manage my migraines?	9	9	9	9	9	-
Why do I still have headaches despite taking painkillers almost every day?	9	9	9	9	9	-
In severe cases, are there surgical options available for treating migraine?	3	3	3	3	3	-
4. Follow-up and Prognosis						
Can migraine be cured?	9	9	9	9	9	-
Will my migraine condition deteriorate over time?	9	9	9	9	9	-
Do migraines elevate the risk of brain-related vascular issues?	9	9	9	9	9	-
How should I monitor and record my migraine symptoms?	9	9	9	9	9	-
5. Special Population Consideration						
What should the elderly pay attention to when suffering from migraine?	9	9	9	9	8	-
I am pregnant/breastfeeding, how should my migraine be treated?	9	9	9	9	9	-
I have menstrual migraines, how should I treat them?	9	9	3[c]	9	9	-
Can children develop migraine, and if so, how does it	9	9	9	9	9	-

affect their health as they grow up?

Each chatbot's score is based on the sum of ratings from three experts. A final rating is considered appropriate when all three experts rate the response as 'good', and poor when any expert rates the response as 'poor'. The proportions of appropriate and poor responses were separately compared across the five LLMs using the $\chi 2$ test.

The numbers in the table represent the total scores assigned by the three reviewers, with 'poor' corresponding to 1 point, 'borderline' to 2 points, and 'good' to 3 points. Specifically, the response was graded as ‘good’ when there were no inaccuracies; ‘borderline’ when there were potential factual inaccuracies but still unlikely to mislead the average patient or cause harm; and ‘poor’ when the response contained unacceptable inaccuracies or inconsistencies that would likely mislead the average patient and cause harm.

[a] The response was considered ‘appropriate’ when all three experts graded it as ‘good’.

[b] The response was considered ‘poor’ when any expert graded it as ‘poor’.

[c] Response from Google Bard for this particular query: I'm not able to help with that, as I'm only a language model.

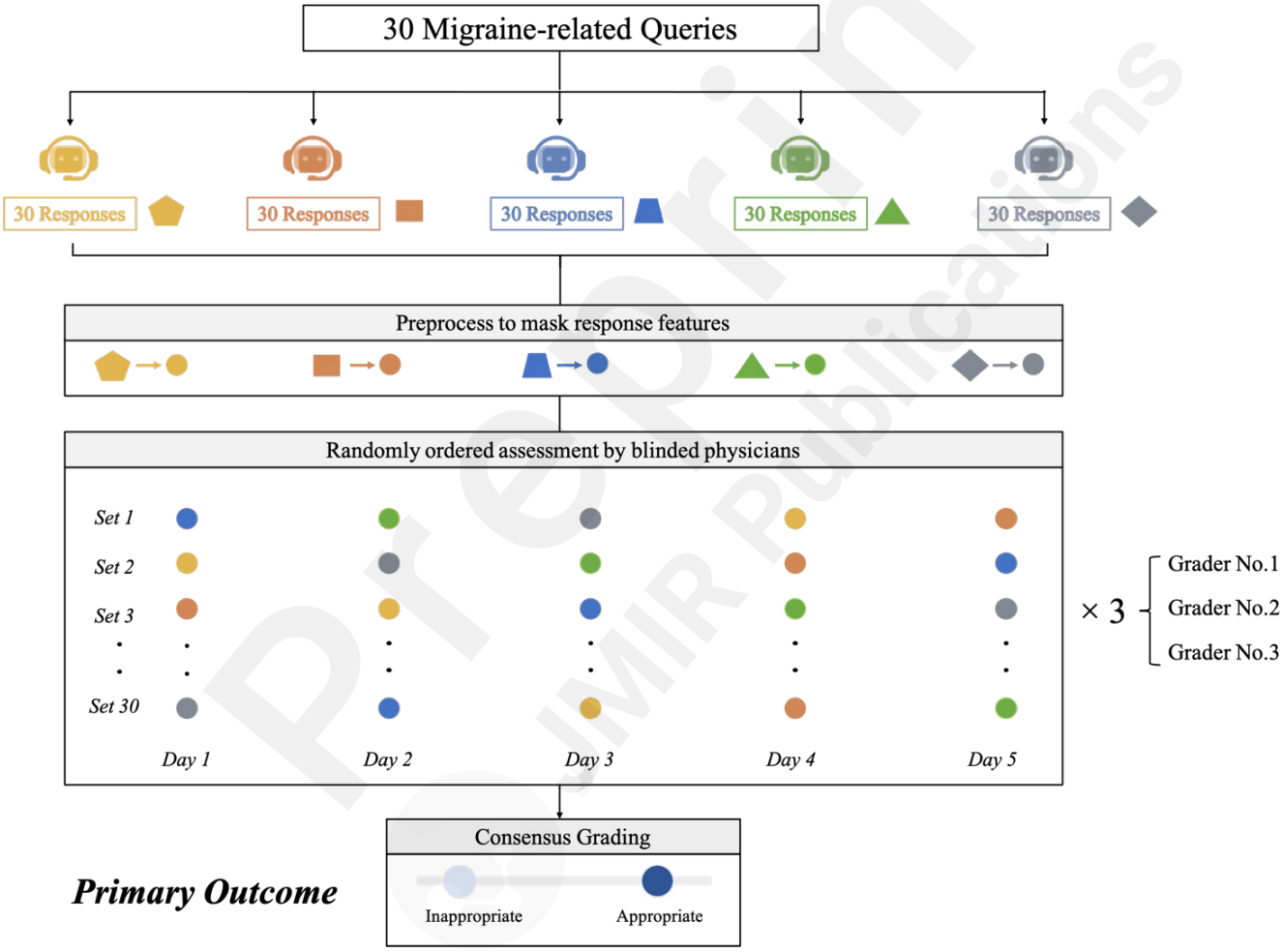

Figure 1. Flowchart of overall study design.

The rating panel comprised three seasoned neurologists, each with at least 15 years of experience in the field. Their primary task was to independently evaluate the accuracy of the LLM chatbots' responses using a three-point scale ('poor,' 'borderline,' and 'good', see annotations for *Table 1*).

Approval from the ethics committee was not required for this study as no patients were involved[8].

Results

ChatGPT-3.5, ChatGPT-4.0, Google Bard, Meta Llama2, and Anthropic Claude2 achieved appropriate response rates of 90%, 96.7%, 83.3%, 90%, and 86.7%, respectively (Pearson's chi-squared test, P=0.481) (*Table 1 and Supplementary Table S1*). Additionally, Google Bard had a 'poor' rating proportion of 6.7%, other LLMs had 3.3% (Pearson's chi-squared test, P=0.961). Notably, none of the LLMs provided an appropriate response to the query, 'In severe cases, are there surgical options available for treating migraine?' This deficiency potentially stems from their limited ability to distinguish between migraine and secondary headaches, a crucial distinction in medical diagnosis. Additionally, the complexity and ongoing debate surrounding surgical interventions for migraine treatment likely contribute to this gap in appropriate advice. Specifically, ChatGPT-3.5 erroneously proposed hemicraniectomy for persistent and severe migraines.

Discussion

Managing migraines over an extended period significantly strains both healthcare systems and patients[2]. In this study, most LLM-responses were accurate and practical, with the majority of responses graded as "good" or "borderline" rather than "poor". Among the five LLMs tested, ChatGPT-4.0 had the highest accuracy, despite that the difference in performance was not statistically significant. These observations suggest that LLMs may function as assistive tools in providing advice, enhancing information acquisition, and offering personalized responses, thereby supporting both patients and physicians. This study represents the first effort to evaluate the capability of state-of-the-art LLMs in educating migraine patients. Although prior studies have primarily focused on migraine diagnostic tests[9], there was no thorough head-to-head comparison of emerging LLMs for migraine patient education. We conducted a rigorous evaluation of five leading

LLMs in addressing common patient queries regarding migraine. The observed variation in performance may be influenced by differences in LLM settings or the limited range of queries tested. Our study benefits from a rigorous study design, incorporating blinding, proper randomization, and an expert review by three neurology specialists. However, it also has limitations. Firstly, the potential generation of false information, such as hallucinations, could lead to patient confusion and even delays in seeking proper medical care. This highlights the need for strategies to mitigate risks, such as improving model bias monitoring and implementing more stringent testing phases before real-world deployment. Secondly, this study is potentially underpowered due to limited sample size, and therefore statistical insignificance should not be viewed as proof of equivalent LLM performance. Additionally, given the continually evolving nature of LLMs, ongoing evaluation with broader datasets is crucial for maintaining the validity of the models and clarifying performance differences between LLMs.

Conclusion

In conclusion, these state-of-the-art LLMs have the potential to accurately respond to common migraine-related queries, which may have implications for generative AI-assisted migraine patient education.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions: Acquisition of data: Linger Li, Pengfei Li, Hongwei Ji, Hongqin Zhao; Analysis and interpretation of data: All authors; Drafting of the manuscript: All authors; Critical revision of the manuscript for important intellectual content: All authors; Statistical analysis: Linger Li, Hongwei Ji; Obtained funding: Hongwei Ji; Study supervision: Hongwei Ji, Hongqin Zhao.

Funding: This study was supported by the National Natural Science Foundation of China (grant number 82103908); the Shandong Provincial Natural Science Foundation (grant number ZR2021QH014); the Shuimu Scholar Program of Tsinghua University; and the National Postdoctoral Innovative Talent Support Program (grant number BX20230189). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflicts of Interest: The authors declare that there is no conflict of interest.

Research ethics and patient consent: Approval from the ethics committee was not required for this study as no patients were involved.

Acknowledgements: None.

Abbreviations: LLMs: large language models.

Reference

1. Collaborators GDaI. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet (London, England)*. 2020;396(10258):1204-1222.
2. Ashina M, Katsarava Z, Do TP, et al. Migraine: epidemiology and systems of care. *Lancet (London, England)*. 2021;397(10283):1485-1495.
3. Feigin VL, Vos T, Alahdab F, et al. Burden of Neurological Disorders Across the US From 1990-2017: A Global Burden of Disease Study. *JAMA neurology*. 2021;78(2):165-176.
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
5. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders, 3rd edition. *Cephalalgia : an international journal of headache*. 2018;38(1):1-211.
6. Ashina M, Buse DC, Ashina H, et al. Migraine: integrated approaches to clinical management and emerging treatments. *Lancet (London, England)*. 2021;397(10283):1505-1518.
7. Robbins MS. Diagnosis and Management of Headache: A Review. *Jama*. 2021;325(18):1874-1885.
8. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770.
9. Cowan RP, Rapoport AM, Blythe J, et al. Diagnostic accuracy of an artificial intelligence online engine in migraine: A multi-center study. *Headache*. 2022;62(7):870-882.

Supplementary Files

Figures

Flowchart of overall study design.