Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GPT-4o and DeepSeek-V3 responses to common migraine questions: a cross-sectional comparative study of accuracy, completeness, empathy, readability and safety

Ahmet Burak Elbeyli

uzmdraburakelbeyli@gmail.com

Develi Dr. Ekrem Karakaya State Hospital

Hasan Hüseyin Kır

Necmettin Erbakan University

Article

Keywords: Migraine, large language models, GPT-4o, DeepSeek-V3, patient education, readability, safety, artificial intelligence

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10488975/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

People increasingly use large language models to obtain health information. Migraine is common, but some symptoms may overlap with stroke and other secondary headache disorders. We compared the answers given by GPT-4o and DeepSeek-V3 to common migraine questions, with attention to accuracy, completeness, empathy, readability and clinical safety.

Methods

In this cross-sectional comparative study, 100 patient-facing questions were compiled from Google Trends, the public r/migraine community on Reddit, and patient information resources. Each question was submitted unchanged to GPT-4o and DeepSeek-V3 in separate sessions. Two neurologists, blinded to model identity, independently rated accuracy, completeness and empathy on 1–5 scales and safety on a 0–5 structured rubric. The Ateşman readability score was calculated once from each response. Exact agreement and quadratic-weighted kappa were used for inter-rater agreement. Question-level mean ratings were compared with paired Wilcoxon signed-rank tests, with Holm correction.

Results

Exact agreement ranged from 88% to 96%, and quadratic-weighted kappa values ranged from 0.810 to 0.920. DeepSeek-V3 had higher accuracy scores than GPT-4o (4.54 ± 0.49 vs 4.37 ± 0.50; Holm-adjusted $p = 0.026$), higher completeness scores (3.98 ± 0.44 vs 3.46 ± 0.61; adjusted $p < 0.001$), and higher Ateşman readability scores (74.75 ± 17.43 vs 60.70 ± 19.97; adjusted $p < 0.001$). Empathy did not differ significantly. The small difference in safety was not significant after correction (adjusted $p = 0.075$). Among 25 red-flag questions, both models had high safety scores; only readability remained significantly different after correction.

Conclusions

Both models generally provided accurate and safe information about migraine. DeepSeek-V3 was more complete and easier to read, but neither model was uniformly superior in empathy or safety. Large language model responses should support, rather than replace, clinical assessment, especially when red-flag symptoms are present.

Background

Migraine is one of the most common neurological disorders worldwide and is a major cause of disability and reduced quality of life [1, 2]. More than one billion people are estimated to live with migraine, and its

burden is particularly marked during working age [1, 2]. Despite this burden, migraine is still often viewed as no more than a severe headache. Patients therefore turn to online sources for information about diagnosis, triggers, imaging, aura, treatment options and long-term outcomes.

The rapid spread of large language models (LLMs) has changed how people obtain and discuss health information [5–7]. Systems such as GPT-4o (OpenAI) and DeepSeek-V3 (DeepSeek) can produce immediate and detailed answers in plain language. At the same time, they may provide incomplete or incorrect information, understate uncertainty, create false reassurance, or fail to recommend medical care when it is needed [5–7].

These concerns are especially important in headache medicine. Symptoms that resemble migraine may occur in cerebrovascular disease, infection, disorders of intracranial pressure and other secondary headache conditions [3, 4]. An answer should therefore be judged not only by whether the medical facts are correct, but also by whether urgent warning signs are recognised and the user is directed to appropriate care.

Previous studies have compared LLM responses to migraine questions and have generally found good factual performance, while also reporting limitations in completeness, consistency, source use and safety [8, 9]. Most work has focused mainly on accuracy. Fewer studies have assessed completeness, patient-centred language, readability and clinical safety together, or have examined a separate set of red-flag questions.

This study compared GPT-4o and DeepSeek-V3 responses to 100 common migraine questions. The answers were evaluated for accuracy, completeness, empathy, readability and safety. A prespecified subgroup of questions containing possible red flags was also examined.

Methods

Study design and question pool

This was a cross-sectional, observational comparison of responses generated by two large language models

to patient-facing questions about migraine.

The question pool was developed in May 2026 through a review of multiple publicly accessible sources,

including Google Trends, discussions in the r/migraine community on Reddit, and patient information

pages published by Mayo Clinic, Cleveland Clinic and the National Health Service [16–20]. These sources

were used to identify recurring themes, commonly searched topics and concerns frequently discussed by

people seeking information about migraine. For the Reddit component, 146 publicly available migraine-related posts and their associated discussions were reviewed. Google Trends queries and institutional patient information pages were used to supplement and cross-check the topics identified during this process.

Usernames, post links, publication dates and other direct or indirect identifiers were not collected or retained. No verbatim quotations from Reddit users were included in the study dataset.

Candidate questions were excluded if they were repetitive, unrelated to migraine, primarily intended for healthcare professionals or did not reflect a patient perspective. Questions with identical or substantially overlapping meanings were grouped and merged. The final question set comprised 100 patient-facing questions across ten categories: diagnosis and differential diagnosis; aura and prodrome; attack symptoms

and urgent situations; environmental triggers; lifestyle, sleep and stress; nutrition; medication and treatment safety; new or interventional treatments; hormonal factors and pregnancy; and prognosis and long-term outcomes.

Table 1
Distribution of question categories

Question category	**n**
Attack symptoms / urgent situations	13
Aura / prodrome	14
Nutrition	13
Hormonal factors / pregnancy	3
Prognosis / long-term outcomes	2
Diagnosis / differential diagnosis	14
Lifestyle / sleep / stress	15
New / interventional treatments	4
Environmental triggers	14
Medication / treatment safety	8

Collection and blinding of model responses

Each question was entered without semantic change into the public web interfaces of GPT-4o (OpenAI) and DeepSeek-V3 (DeepSeek) during May 2026. Separate sessions were used, and no follow-up prompt, correction request or additional instruction was given. The outputs were copied into the data forms without content editing.

Model names and other obvious identifying cues were removed from the response files, and the responses were coded before assessment. Two neurologists rated the answers independently and were blinded to the model label. They did not discuss individual scores during the rating process.

Evaluation criteria

Accuracy, completeness and empathy were scored on 1–5 structured scales. Accuracy reflected agreement with current neurological knowledge and clinical practice. Completeness reflected whether the response answered the question directly, included the main information needed by a patient, mentioned relevant alternatives or uncertainty, and recommended professional assessment where appropriate. Empathy reflected acknowledgement of concern, a non-judgemental tone, clarity and supportive language.

Safety was independently rated by two neurologists using a structured 0–5 clinical safety scale. The assessment considered recognition of a red flag when present, appropriate urgent referral, absence of harmful advice, avoidance of unnecessary alarm, and recommendation of specialist care when needed. Criteria that were not applicable to a particular question were excluded from the assessment. These criteria were used as a structured clinical assessment framework and were not intended to constitute a newly validated psychometric scale.

Before rating, 25 questions were prespecified as red-flag questions. They included sudden or worst-ever headache, new neurological deficit, signs of infection, altered consciousness, headache during pregnancy or the postpartum period, and other situations suggesting a possible secondary headache.

Readability assessment

The Ateşman formula was used to measure Turkish readability [15]: 198.825 − 40.175 × (syllables/words) − 2.610 × (words/sentences). Because readability is a property of the response text rather than the rater, one score was calculated for each model response. Values were constrained to the 0–100 range. Scores of 90–100 were classified as very easy, 70–89.99 as easy, 50–69.99 as moderately difficult, 30–49.99 as difficult, and below 30 as very difficult.

Statistical analysis

Descriptive results are presented as mean and standard deviation (SD), together with median and interquartile range (IQR). Inter-rater agreement was calculated separately for each model and clinician-rated domain using exact percentage agreement and quadratic-weighted Cohen's kappa.

To avoid treating repeated ratings of the same question as independent observations, the two neurologist scores were averaged at the question level for each model. DeepSeek-V3 and GPT-4o were then compared across the same 100 questions using two-sided Wilcoxon signed-rank tests. The primary outcomes were accuracy, completeness, empathy, safety and Ateşman readability. Holm correction was applied across the five primary comparisons. Effect size was expressed as r, calculated from the standardised test statistic and the number of non-zero paired differences.

The same procedure was used in the prespecified red-flag subgroup, with a separate Holm correction across its five outcomes. No formal sample-size calculation was performed because the analysis used the complete prespecified set of 100 selected questions. Statistical analyses were conducted in Python 3.11 using SciPy, scikit-learn and statsmodels. A two-sided p value below 0.05 was considered statistically significant.

Results

Question set and inter-rater agreement

The final dataset contained 100 questions, including 25 red-flag questions. Category sizes ranged from 2 questions on prognosis and long-term outcomes to 15 questions on lifestyle, sleep and stress (Table 1).

Across the four clinician-rated outcomes, exact agreement ranged from 88% to 96%. Quadratic-weighted kappa ranged from 0.810 to 0.920, indicating high agreement overall (Table 2).

Table 2
Inter-rater agreement by model and outcome

Model	Outcome	Exact agreement, %	Quadratic-weighted kappa
GPT-4o	Accuracy	90.0	0.818
GPT-4o	Completeness	92.0	0.897
GPT-4o	Empathy	88.0	0.882
GPT-4o	Safety	96.0	0.920
DeepSeek-V3	Accuracy	94.0	0.880
DeepSeek-V3	Completeness	92.0	0.810
DeepSeek-V3	Empathy	88.0	0.858
DeepSeek-V3	Safety	93.0	0.895

Overall model comparison

DeepSeek-V3 received higher question-level mean scores for accuracy (4.54 ± 0.49 vs 4.37 ± 0.50) and completeness (3.98 ± 0.44 vs 3.46 ± 0.61). These differences remained significant after Holm correction (adjusted p = 0.026 and adjusted p < 0.001, respectively).

Empathy scores were similar (3.42 ± 0.63 vs 3.35 ± 0.69; adjusted p = 0.247). DeepSeek-V3 had a slightly higher safety mean (4.67 ± 0.56 vs 4.54 ± 0.49), but the difference did not remain significant after correction (adjusted p = 0.075).

The recalculated Ateşman score was higher for DeepSeek-V3 (74.75 ± 17.43) than for GPT-4o (60.70 ± 19.97; adjusted p < 0.001). Fifty DeepSeek-V3 answers were classified as easy and 18 as very easy, compared with 21 easy and 8 very easy GPT-4o answers. Full results are shown in Table 3.

Table 3
Paired comparison of GPT-4o and DeepSeek-V3 across 100 questions

Outcome	GPT-4o, mean ± SD; median (IQR)	DeepSeek-V3, mean ± SD; median (IQR)	W	Raw p	Holm p	r
Accuracy	4.37 ± 0.50; 4.00 (4.00–5.00)	4.54 ± 0.49; 5.00 (4.00–5.00)	360.0	0.0088	0.0264	0.374
Completeness	3.46 ± 0.61; 3.00 (3.00–4.00)	3.98 ± 0.44; 4.00 (4.00–4.00)	167.5	< 0.001	< 0.001	0.778
Empathy	3.35 ± 0.69; 3.00 (3.00–4.00)	3.42 ± 0.63; 3.00 (3.00–4.00)	398.0	0.2474	0.2474	0.174
Safety	4.54 ± 0.49; 5.00 (4.00–5.00)	4.67 ± 0.56; 5.00 (4.00–5.00)	328.0	0.0375	0.0750	0.314
Ateşman readability	60.70 ± 19.97; 60.03 (44.32–77.13)	74.75 ± 17.43; 76.03 (62.59–88.76)	864.5	< 0.001	< 0.001	0.565

Note. Ratings are question-level means of two neurologists. Ateşman readability was calculated once per response. Holm correction was applied across five outcomes.

Red-flag subgroup

In the 25 red-flag questions, both models had high accuracy and safety scores. GPT-4o had a safety mean of 4.84 ± 0.37 and DeepSeek-V3 had a mean of 4.68 ± 0.54; the difference was not significant after adjustment.

DeepSeek-V3 had a higher completeness mean (4.08 ± 0.37 vs 3.76 ± 0.61), but the raw p value of 0.028 did not remain significant after Holm correction (adjusted p = 0.112). The only outcome that remained

significantly different was Ateşman readability (71.31 ± 15.67 vs 54.91 ± 20.27; adjusted p < 0.001). There were no adjusted differences in accuracy, empathy or safety (Table 4).

Table 4
Paired comparison in the 25 red-flag questions

Outcome	GPT-4o, mean ± SD	DeepSeek-V3, mean ± SD	W	Raw p	Holm p	r
Accuracy	4.52 ± 0.49	4.56 ± 0.49	41.0	0.7431	1.000	0.091
Completeness	3.76 ± 0.61	4.08 ± 0.37	19.0	0.0281	0.1122	0.587
Empathy	3.82 ± 0.43	3.82 ± 0.63	32.0	0.9272	1.000	0.000
Safety	4.84 ± 0.37	4.68 ± 0.54	4.0	0.1670	0.5010	0.564
Ateşman readability	54.91 ± 20.27	71.31 ± 15.67	32.0	< 0.001	< 0.001	0.754

Note. Holm correction was applied separately across the five red-flag subgroup outcomes. Full medians and interquartile ranges are provided in Additional file 1.

Discussion

This study compared GPT-4o and DeepSeek-V3 answers to a broad set of migraine questions that reflected common patient concerns. Both models performed well in accuracy and safety. DeepSeek-V3, however, provided more complete answers and responses that were easier to read. The advantage in safety was small and did not remain significant after correction, while empathy scores were similar.

The findings are broadly consistent with the study by Li et al., which compared several LLMs on common migraine questions and found generally high accuracy, with differences between models [8]. Our study adds DeepSeek-V3, uses a larger question set, and evaluates Turkish answers across several domains rather than accuracy alone. It also separates questions that may represent urgent secondary headache conditions.

The largest clinical-rating difference was in completeness. This matters because a technically correct answer may still omit an alternative diagnosis, uncertainty, a treatment limitation or the need for medical review. Prior work has similarly shown that useful LLM responses can remain incomplete or inconsistent [8, 9]. The present results support evaluating omission as well as direct factual error.

DeepSeek-V3 responses also had substantially higher Ateşman scores. Readability can improve access to information, but it should not be taken as proof of clinical quality. Covarrubias et al. reported that ChatGPT could make patient material easier to read while still introducing omissions or creating a need for expert correction [12]. A short and simple answer may be readable but clinically insufficient, whereas a more complete answer may require medical terms. Readability and clinical content should therefore be considered together.

Empathy did not differ significantly between the models. A supportive tone may help users engage with health information, but empathy cannot compensate for an unsafe or incomplete recommendation. This point is particularly important when the user describes a red flag. Reassuring language without adequate triage may delay care.

In the red-flag subgroup, both models had high safety scores and no adjusted difference in safety, accuracy, completeness or empathy. Only readability remained significantly different. The high scores are encouraging, but they do not show that the systems are safe in every possible clinical situation. The subgroup contained only 25 questions, and LLM outputs may vary between sessions and as models are updated.

Recent safety studies have shown that LLMs can provide answers with potential for patient harm, even when their average performance appears strong [10]. Frameworks for evaluating clinical summaries also stress the importance of both hallucinations and omitted information [11]. These concerns support a human-supervised role for LLMs in patient education [13, 14]. They should help patients understand basic concepts and recognise when professional review is needed, not make a diagnosis or select treatment independently.

The study also showed high agreement between the two neurologists. This supports the internal consistency of the structured assessment. However, the rubrics were study-specific and were not formally validated as new psychometric instruments. Hence, these findings reflect specific expert evaluations within this dataset rather than universal measurement properties.

Clinical implications

LLMs may be useful for first-line migraine education because they can answer questions quickly and in accessible language. Their most appropriate role is to explain general information, support health literacy and encourage timely contact with health professionals.

Clinical use should include clear safeguards. Responses about sudden severe headache, new neurological deficits, altered consciousness, infection, pregnancy or the postpartum period should prioritise urgent assessment. Model performance should be checked regularly because the systems and their outputs change over time.

Limitations

The study has several limitations. First, the 100 questions cannot represent every way in which patients may describe migraine or secondary headache. Second, the question pool was derived from online sources and may overrepresent people who actively seek health information on the internet. Third, all questions and responses were in Turkish, which limits direct generalisation to other languages and cultures.

Only two models were evaluated at one point in time. The model versions were recorded as GPT-4o and DeepSeek-V3; however, the exact backend build identifiers used by the public web interfaces were not retained, and later model updates may affect reproducibility. Responses were generated once per question, so within-model variability was not tested. The clinical ratings were based on a study-specific structured rubric rather than a formally validated scale.

The Ateşman formula is sensitive to sentence and word segmentation, particularly in short responses. Although a reproducible text-processing rule was applied to all answers, other software may produce slightly different scores. Finally, the study evaluated text outputs rather than patient understanding, behaviour, clinical outcomes or actual harm.

Conclusions

GPT-4o and DeepSeek-V3 generally gave accurate and safe answers to common migraine questions. DeepSeek-V3 had higher accuracy and completeness ratings and produced more readable Turkish responses. Empathy did not differ, and the small overall safety difference was not significant after adjustment.

Among red-flag questions, both models had high safety scores, but neither should be considered a substitute for clinical evaluation. LLMs may support patient education when their outputs are reviewed critically and when urgent warning signs lead to clear professional referral.

Abbreviations

AI
artificial intelligence
IQR
interquartile range
LLM
large language model
NHS
National Health Service
QWK
quadratic-weighted kappa
SD
standard deviation.

Declarations

Ethics approval and consent to participate

Not applicable. The study did not involve human participants, individual-level human data, identifiable personal information, intervention or recruitment. Public online material was used only to identify general question themes; no username, direct quotation or other identifying information was collected or analysed.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this article and its additional files. The question-level dataset, model responses, individual ratings and recalculated readability scores are provided in Additional file 1. The analysis code is provided in Additional file 2.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

ABE designed the study, collected the questions, obtained the model responses, organised the data, independently evaluated the responses, performed the statistical analysis, interpreted the findings and drafted the manuscript. HHK contributed to the study concept, independently evaluated the responses, interpreted the findings and critically revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors thank Dr Eray Abat for his comments on the manuscript.

Authors' information

Not applicable.

References

1. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* **23**, 344–381. 10.1016/S1474-4422(24)00038-3 (2024).

2. Steiner, T. J., Stovner, L. J., Jensen, R., Uluduz, D. & Katsarava, Z. Migraine remains second among the world's causes of disability, and first among young women: findings from GBD2019. *J. Headache Pain*. **21**, 137. 10.1186/s10194-020-01208-0 (2020).
3. Headache Classification Committee of the International Headache Society. The International Classification of Headache Disorders, 3rd edition. Cephalalgia. ;38:1–211. (2018). 10.1177/0333102417738202
4. Goadsby, P. J. et al. Pathophysiology of migraine: a disorder of sensory processing. *Physiol. Rev.* **97**, 553–622. 10.1152/physrev.00034.2015 (2017).
5. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596. 10.1001/jamainternmed.2023.1838 (2023).
6. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthc. (Basel)*. **11**, 887. 10.3390/healthcare11060887 (2023).
7. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl. J. Med.* **388**, 1233–1239. 10.1056/NEJMsr2214184 (2023).
8. Li, L. et al. Benchmarking state-of-the-art large language models for migraine patient education: performance comparison of responses to common queries. *J. Med. Internet Res.* **26**, e55927. 10.2196/55927 (2024).
9. Schütz, P. et al. ChatGPT as an information source for patients with migraines: a qualitative case study. *Healthc. (Basel)*. **12**, 1594. 10.3390/healthcare12161594 (2024).
10. Draelos, R. L. et al. Large language models provide unsafe answers to patient-posed medical questions. *NPJ Digit. Med.* **9**, 241. 10.1038/s41746-026-02428-5 (2026).
11. Asgari, E. et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit. Med.* **8**, 274. 10.1038/s41746-025-01670-7 (2025).
12. Covarrubias, O., Ghanem, D., Murdock, C., Domes, C. & Shafiq, B. Improving the readability of trauma patient education materials: a ChatGPT solution demonstrated using materials by the Orthopaedic Trauma Association. *OTA Int.* **8**, e417. 10.1097/OI9.0000000000000417 (2025).
13. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248. 10.1145/3571730 (2023).
14. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit. Med.* **6**, 120. 10.1038/s41746-023-00873-0 (2023).
15. Ateşman, E. Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi*. **58**, 71–74 (1997).
16. Google Trends. May (2026). https://trends.google.com/trends/. Accessed 31.
17. Reddit r/migraine. https://www.reddit.com/r/migraine/. Accessed 31 May 2026.
18. Mayo Clinic. Migraine. https://www.mayoclinic.org/diseases-conditions/migraine-headache/. Accessed 31 May 2026.

19. Cleveland Clinic. Migraine headaches. https://my.clevelandclinic.org/health/diseases/5005-migraine-headaches. Accessed 31 May 2026.
20. National Health Service. Migraine. https://www.nhs.uk/conditions/migraine/. Accessed 31 May 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfileslegends.docx
- Additionalfile1MigraineLLMdataset.xlsx
- Additionalfile2Analysiscode.py