Neurology Asia 2026; 31(3) : 1055 – 1063

Performance comparison of large language models in interpreting clinical guidelines for migraine prevention: A multidimensional analysis

*[1,2]Li Xu, *[1,2]Xu Qiu, *[6]Jiayi Deng, [1,2]Chengqi Dong, [1,2]Dong Liang, [1,3]Keyang Wu, [4,5]Xiaoxue Dong, [1,2]Tao Mei, [1,2]Shi Chen, [1,2]Yali Wu, [1,2,3]Yuan Cheng, [1,2,3]Jianliang Sun, [1,2]Liang Yu, [1,2]Hanbin Wang, [1,2]Qinghua Li

*L Xu, X Qiu, JY Deng contributed equally to this work and are co-first authors

[1]The Fourth School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou First People's Hospital, Hangzhou, China; [2]Department of Pain, The Affiliated Hangzhou First People's Hospital, Westlake University School of Medicine, Hangzhou, China; [3]Department of Anesthesiology, the Affiliated Hangzhou First People's Hospital, Westlake University School of Medicine, Hangzhou, China; [4]National Neuroscience Institute of Singapore, Singapore; [5]Department of Neurology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [6]Department of Pain, Wuxi Xishan People's Hospital, Wuxi, China.

Abstract

Background & Objective: Large language models (LLMs) such as DeepSeek-R1, Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert are increasingly applied in medical contexts, yet their reliability in evidence-based clinical domains like migraine prophylaxis remains uncertain. This study aimed to compare the performance of four leading LLMs in interpreting and applying the International Headache Society's global practice recommendations for preventive pharmacological treatment of migraine. *Methods:* Sixteen standardized clinical scenario questions derived from the IHS guideline were presented identically to each model. Responses were evaluated by blinded expert raters across five dimensions—Accuracy, Overconclusiveness, Supplementary Value, Incompleteness, and Readability—using a 10-point Likert scale. Readability was further analyzed using composite indices from readabilityformulas.com. *Results:* No significant inter-model differences were observed in Accuracy (P = 0.856), Overconclusiveness (P = 0.400), or Incompleteness (P = 0.531). However, DeepSeek-R1 provided significantly more Supplementary Information than Gemini-2.5 Pro (P = 0.010) and Grok-4 Expert (P = 0.030). Readability analysis further revealed substantial variation across models (P < 0.001), with DeepSeek-R1 generating the most accessible outputs.
Conclusion: While all four models exhibited comparable adherence to guideline-based content, DeepSeek-R1 demonstrated superior performance in supplementary informational value and readability. These findings highlight the importance of evaluating LLMs not only for accuracy but also for their capacity to enhance clinical communication and decision support.

Keywords: Large language models, migraine, artificial intelligence

INTRODUCTION

Large Language Models (LLMs), as a revolutionary breakthrough in artificial intelligence (AI), have become deeply integrated into many aspects of daily life and are penetrating the highly demanding field of medicine at an unprecedented pace. From assisting in clinical diagnosis and interpreting medical images to accelerating drug development and drafting scientific papers, LLMs have demonstrated immense potential in streamlining information workflows, consolidating vast amounts of literature, and providing decision support.[1-3]

Address correspondence to: Qinghua Li, Hanbin Wang, Liang Yu, all of Department of Pain, the Affiliated Hangzhou First People's Hospital, Westlake University School of Medicine, No. 261, Huansha Road, Shangcheng district, Hangzhou 310006, China. Emails: QH Li, 13858149400@163.com; HB Wang, wanghanbin@hospital.westlake.edu.cn; L Yu, yuliang@hospital.westlake.edu.cn

Date of Submission: 24 February 2026; Date of Acceptance: 21 March 2026

https://doi.org/10.54029/2026css

However, when these general-purpose models are applied to specific clinical scenarios that demand precise, evidence-based decisions, their reliability, accuracy, and safety become critical issues requiring thorough validation.

Migraine is precisely such a challenging clinical domain. As a highly prevalent and disabling neurological disorder, migraine imposes a heavy physiological, psychological, and economic burden on hundreds of millions of patients worldwide.[4] Epidemiological data show that it consistently ranks among the leading causes of global disease burden, particularly affecting the work and quality of life of young and middle-aged adults.[5] For patients with moderate to high-frequency attacks, preventive pharmacotherapy is the core strategy for managing migraine, reducing attack frequency, and lowering the risk of disability. In clinical practice, however, the selection of preventive medication is exceedingly complex, requiring comprehensive consideration of efficacy, side effects, patient comorbidities, and individual needs.

To standardize and guide clinical decision-making, the International Headache Society (IHS) has developed and published the "Global Practice Recommendations for Preventive Pharmacological Treatment of Migraine".[6] This guideline, based on the best available evidence and expert consensus, provides a clear and practical framework for medication selection and application for clinicians globally, especially for those in resource-limited settings. It serves not only as a cornerstone for standardized migraine treatment but also sets a high bar for the accuracy required of such clinical support tools.

Against this backdrop, combining the rapid information-processing capabilities of LLMs with strict clinical guidelines gives rise to new opportunities and challenges. In theory, LLMs could help physicians quickly retrieve, understand, and explain complex guideline recommendations to patients, thereby optimizing clinical workflows. On the other hand, different models vary substantially in architecture, training data, conversational strategies, and internal reasoning mechanisms. Representative examples include DeepSeek-R1, Google's Gemini family, OpenAI's GPT-5, and xAI's Grok models — each with distinct strengths and safety trade-offs. The extent to which LLM-generated responses align with the gold standard of clinical practice directly impacts their utility and safety. It is therefore necessary to systematically evaluate the performance of cutting-edge LLMs in interpreting and applying clinical guidelines, establishing a foundation for their safe and effective deployment in the medical field.

Therefore, this study conducts a systematic, multi-dimensional comparison of DeepSeek-R1, Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert on a standardized question set derived from the IHS preventive treatment recommendations. We aim to determine whether models differ in terms of Accuracy, Over conclusiveness, Supplementary Value, Incompleteness, and Readability, thereby providing evidence-based insights for the selection and application of large language models in the clinical medication management of migraine. By systematically comparing these key dimensions, this study seeks to support clinicians in leveraging the distinctive strengths of different LLMs while implementing necessary safeguards to minimize potential risks.

METHODS

Study design

This is an LLM-based content-analysis, cross-sectional paired-comparison study. The study utilized the 16 specific clinical questions explicitly formulated in the 2024 IHS Global Practice Recommendations (hereafter referred to as the "Migraine Guideline"). These questions represent the consensus-defined core decision points in migraine management. Using these pre-validated questions ensured that the study's scope directly mirrors the authoritative clinical framework. We presented each identically to four large language models (DeepSeek-R1, Gemini-2.5 Pro, Chat-GPT-5 Thinking, and Grok-4 Expert) and systematically evaluated their responses across five key dimensions: Accuracy, Overconclusiveness, Supplementary Value, Incompleteness, and Readability.

Data collection

The standardized set of 16 clinical scenario questions was submitted to four large language models—DeepSeek-R1, ChatGPT-5 (with thinking mode enabled), Gemini-2.5 Pro, and Grok-4 Expert—via their publicly accessible interfaces between September 10 and 30, 2025. To ensure consistency, all non-specified parameters (e.g., temperature, decoding settings) were kept at their default values, and system

prompts were not standardized. Browsing mode and safety filters were deactivated for all models. No specific version numbers were recorded beyond the model names provided. To mitigate data retention bias, each question was asked only once in a separate chat session, and after each session, the conversation and browser history were cleared.

Following the initial round of queries, two additional rounds were conducted to assess the variability of the AI models' responses. The full set of responses from all three rounds is provided in Supplementary S1–S5. The comparison included a quantitative assessment of concordance and divergence across several domains: accuracy of recommended treatments, tendency toward over-conclusiveness, provision of supplementary information, and extent of incompleteness relative to the reference.

Evaluation process

A total of three senior clinical experts independently evaluated the responses, all of whom possess extensive clinical experience, including one neurologist and two pain specialists. They evaluated responses in a blinded fashion: model identity and presentation order were removed and items randomized. Raters received standardized training on the scoring protocol with example items to improve consistency.

For each response, raters used a 10-point Likert scale (1–10) to score four dimensions, with the final score for each question and dimension calculated as the arithmetic mean of the three independent ratings:

Accuracy: How closely the answer aligns with IHS guideline recommendations (1 = completely incorrect/misleading; 10 = fully accurate).

Overconclusiveness: Tendency to make overly definitive assertions beyond the evidence (1 = extremely overconclusive; 10 = very cautious).
Supplementary Value: Value of additional content beyond core guideline recommendations (e.g., monitoring tips, alternative options, evidence caveats) (1 = none; 10 = exceptional).

Incompleteness: Degree of missing critical information (e.g., contraindications, dose ranges, special-population caveats) (1 = very incomplete; 10 = very complete). These five dimensions were selected to balance clinical reliability (Accuracy and Incompleteness), safety (Overconclusiveness), and practical utility (Supplementary Value and Readability), following established paradigms in recent AI-medical evaluation literature.

Inter-rater reliability was quantified using ICC[2,1] across all four models combined (n=64 items per dimension). Overall ICC values were: Accuracy = 0.626, Overconclusiveness = 0.715, Supplementary Value = 0.631, Incompleteness = 0.674 (see Supplementary Table S1).

Readability analysis

Readability was assessed using *readabilityformulas.com*, applying the Automated Readability Index (ARI), Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Coleman–Liau Index (CLI), the SMOG Index, and the FORCAST Readability Formula. Each formula captures different aspects of textual complexity: FRE evaluates readability based on the number of words, sentences, and syllables; FKGL estimates the U.S. school grade level required to comprehend the text; GFI considers average sentence length and the proportion of complex words; CLI estimates educational level based on counts of letters, words, and sentences, independent of syllables; and the SMOG Index predicts the years of education needed to understand the text with ease. ARI and FORCAST provide alternative grade-level estimates, with ARI relying on characters per word and words per sentence, and FORCAST optimized for technical materials without using sentence length. These measures emphasize different perspectives of reading difficulty—some emphasize the challenge of longer sentences, while others highlight technical vocabulary.

To harmonize these differences and allow standardized comparison, we used the built-in Average Reading Level Calc (ARLCalc) tool of readabilityformulas.com, which aggregates multiple readability indices into a single "average reading grade" score. This composite metric is referred to as the Average Reading Level (ARL) hereafter. ARLCalc serves as a normalization and consensus method rather than a novel readability formula, enabling more accurate cross-model comparisons.

Statistical analysis

Mean scores per model and dimension were compared via one-way ANOVA and two-way ANOVA; Tukey HSD used for pairwise post-hoc comparisons when ANOVA was significant. Alpha = 0.05. Means ± SD and p-values reported.

RESULTS

In this study, we evaluated the performance of four large language models—DeepSeek-R1, Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert—by inputting 16 clinical scenario questions derived from the Migraine Guideline. Each question was asked only once in a separate chat session to assess the models' responses. The outputs were evaluated using a 10-point Likert scale across four dimensions: Accuracy, Over conclusiveness, Supplementary Information, and Incompleteness. This approach allowed for a comprehensive comparison of each model's alignment with established clinical guidelines. The responses generated by all evaluated models were systematically documented in Supplementary S1–S4.

As shown in Figure 1, the performance of the four models was evaluated across four distinct metrics. For the Accuracy metric, no statistically significant differences were observed among the models (F = 0.257, P = 0.856). Similarly, assessments of Over-conclusiveness revealed no significant inter-model variations (F = 0.999, P = 0.400). The analysis of Incompleteness also showed non-significant differences across the four models (F = 0.743, P = 0.531).

In contrast, a statistically significant difference was detected in the Supplementary information category (F = 4.247, P = 0.009). A subsequent Tukey's post-hoc test specified that DeepSeek-R1 provided significantly more supplementary information than both Gemini-2.5 Pro (P = 0.010) and Grok-4 Expert (P = 0.030). The mean differences were 0.71 (95% CI: 0.13–1.28) and 0.63 (95% CI: 0.05–1.20), respectively. In summary, while the four AI models performed comparably on three of the four dimensions, DeepSeek-R1 showed a statistically significant outperformance in the provision of supplementary content.

This statistically significant outperformance of DeepSeek-R1 in the supplementary information category is further detailed by the score distribution in Figure 5. Across the four evaluation categories, all models exhibited mean scores above 6, and none of the responses attained the maximum score of 10. The analysis reveals that 70.6% of DeepSeek-R1's responses achieved a high rating (scores ≥ 8). This proportion is notably higher than that of both Gemini-2.5 Pro (37.5%) and Grok-4 Expert (62.5%), a finding that supports the significant differences observed in their respective mean scores (P = 0.010 and P = 0.030, respectively).

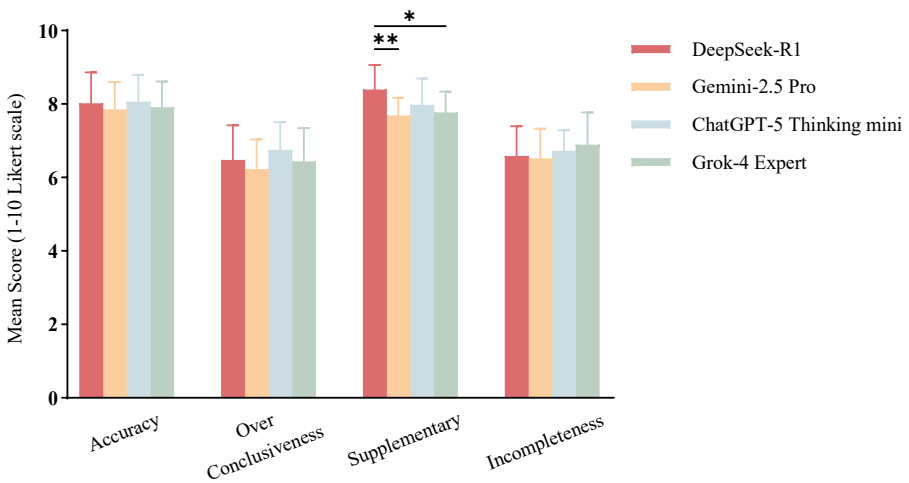

Figure 1. Comparative accuracy, over-conclusiveness, supplementary information, and incompleteness of LLMs. Key findings: 1. Accuracy: non-significant (P = 0.856) 2. Over-conclusiveness: non-significant (P = 0.400) 3. Supplementary: significant (P = 0.009), with DeepSeek-R1 scoring higher than Gemini-2.5 Pro (P = 0.010) and Grok-4 Expert (P = 0.030) 4. Incompleteness: non-significant (P = 0.531)

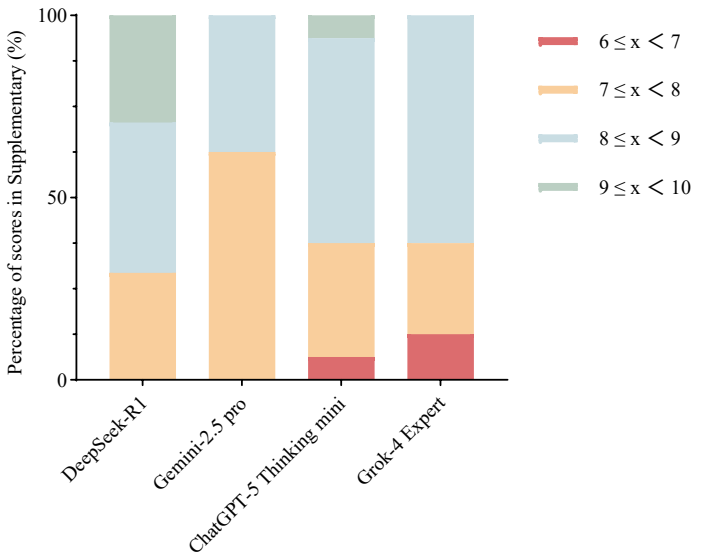

Figure 2. Proportion of guideline-compliant responses in Supplementary by LLM. The 100% stacked bar chart illustrates the percentage of scores from each model that fall into one of four score bins ($6 \leq x < 7$, $7 \leq x < 8$, $8 \leq x < 9$, and $9 \leq x \leq 10$), where 'x' represents the individual evaluation score.

Readability

Analysis of ARL revealed significant differences among the models ($P < 0.001$, $F = 27.23$). As shown in Figure 3, where a lower ARL signifies easier reading, the output from DeepSeek-R1 was found to be significantly more readable than the text generated by Gemini-2.5 Pro ($P < 0.001$, Mean diff = -2.38, 95% CI [-3.27, -1.50]), ChatGPT-5 Thinking ($P < 0.001$, Mean

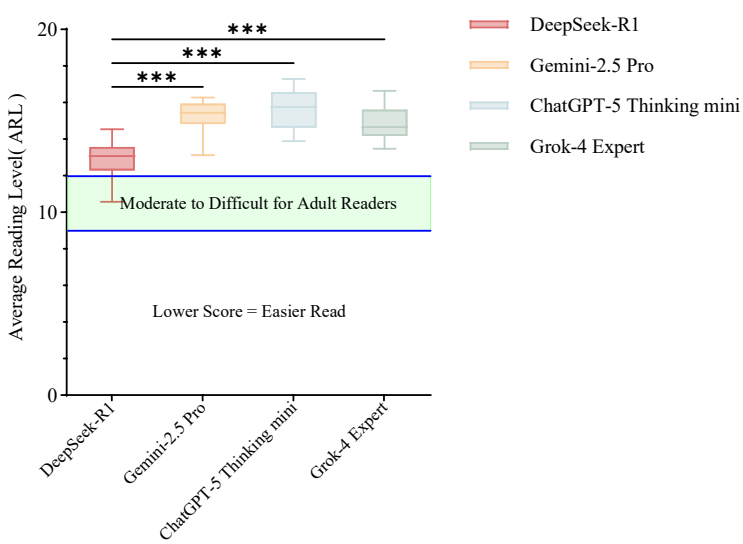

Figure 3. Readability of texts generated by large language models. The Average Reading Level (ARL) of outputs from four models is compared using box plots, where a lower score indicates easier readability. DeepSeek-R1 generated significantly more readable text (lower ARL) than Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert. The median readability of DeepSeek-R1's output corresponds to a "Moderate to Difficult" level for adults (shaded area), whereas the other models produced text of higher complexity. The central line in each box represents the median, with the box boundaries indicating the interquartile range. Significance levels are denoted by asterisks (*P < 0.05, **P < 0.01, ***P < 0.001).

diff = -2.79, 95% CI [-3.68, -1.90]), and Grok-4 Expert (P < 0.001, Mean diff = -1.91, 95% CI [-2.80, -1.03]). The median readability score for DeepSeek-R1 placed its output within the "Moderate to Difficult" range for adult readers, whereas the outputs from the other three models were categorized as being of higher complexity.

FRE assessment, a dimension of the overall ARL analysis, yielded congruent results (P < 0.001, F =15.90). As shown in Fig. 4, DeepSeek-R1 exhibited a significantly higher FRE score than Gemini-2.5 Pro (P < 0.001, Mean diff = 14.19, 95% CI [7.03, 21.35]), ChatGPT-5 Thinking (P < 0.001, Mean diff = 16.38, 95% CI [9.22, 23.53]), and Grok-4 Expert (P < 0.001, Mean diff = 14.94, 95% CI [7.78, 22.10]), classifying it as 'Difficult to read'. The outputs from the other three models, with median scores below 30, were rated as 'Confusing to read'.

As shown in Figure 5, DeepSeek-R1 consistently produced texts with lower median grade levels across the ARI, GFI, FKGL, CLI, SMOG and FORCAST indices compared to the other models, which exhibited medians ranging from approximately 10 to 15. These differences were significant (P < 0.05) for all comparisons in these five indices, indicating that DeepSeek-R1 outputs are substantially easier to read based on these metrics.

DISCUSSION

Despite the growing enthusiasm for integrating LLMs into clinical workflows[7-9], concerns remain regarding their factual reliability and interpretive precision when applied to evidence-based medicine.[10,11] To address these gaps, we introduce a reproducible evaluation paradigm—comprising a standardized, guideline-derived question set, blinded parallel scoring by trained expert raters, multidimensional metrics (accuracy, overconclusiveness, supplementary value, incompleteness), and objective readability indices—that maximizes internal validity and enables cross-study comparability, and we apply this paradigm here in one of the first systematic cross-model evaluations of LLMs against guideline-based migraine prophylaxis recommendations.

The comparable performance in accuracy

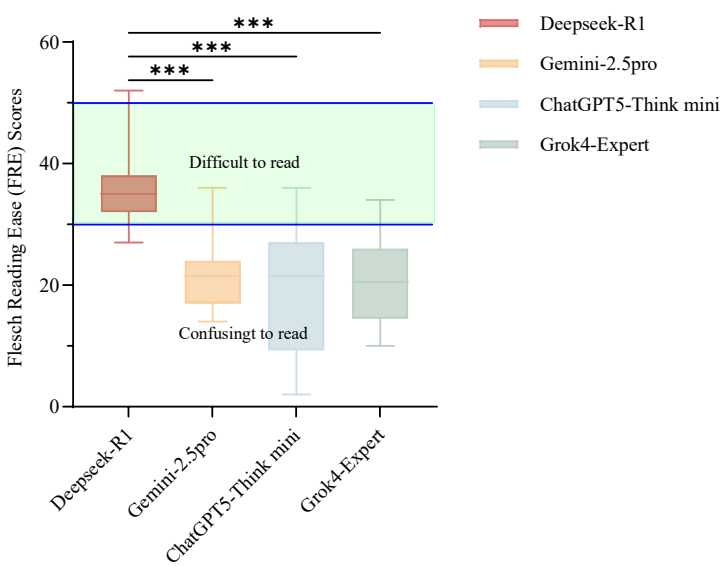

Figure 4. Comparison of Flesch Reading Ease (FRE) scores for texts generated by four large language models. Box plots show the distribution of FRE scores, where higher values indicate easier readability. Outputs from DeepSeek-R1 achieved a significantly higher FRE score than those from Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert. Based on the standard FRE scale, the median score for DeepSeek-R1 falls into the 'Difficult to read' range (shaded area; FRE 30-50), whereas the median scores of the other three models are categorized as 'Confusing to read' (FRE < 30). The central line in each box represents the median, with the box boundaries indicating the interquartile range. Significance levels are denoted by asterisks (*P < 0.05, **P < 0.01, ***P < 0.001).

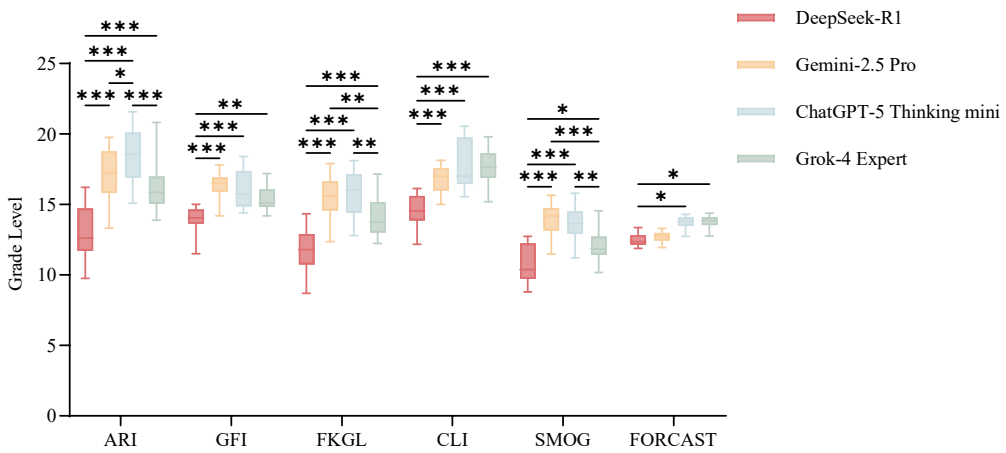

Figure 5. Comparison of readability grade levels for texts generated by four large language models across multiple indices: the Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch–Kincaid Grade Level (FKGL), Coleman–Liau Index (CLI), the SMOG Index, and the FORCAST Readability Formula. Box plots show the distribution of grade levels, where lower values indicate easier readability. Outputs from DeepSeek-R1 achieved significantly lower grade levels than those from Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert across all indices. The central line in each box represents the median, with the box boundaries indicating the interquartile range. Significance levels are denoted by asterisks (*P < 0.05, **P < 0.01, ***P < 0.001).

suggests that recent-generation LLMs have achieved a high baseline alignment with clinical recommendations, possibly reflecting advances in model architecture, reinforcement learning, and fine-tuning strategies using domain-specific corpora.[12] However, accuracy alone does not guarantee safe clinical applicability. Overconclusiveness—the tendency to assert beyond available evidence—has been highlighted as a key risk factor in clinical AI deployment.[13-15] Our analysis found no significant differences across models in this dimension, which may indicate improved calibration mechanisms in current systems. Nevertheless, qualitative inspection revealed that all models occasionally introduced speculative elements, underscoring the need for continued human oversight in clinical decision-making.[16,17]

DeepSeek-R1's stronger provision of supplementary content may increase the practical value of guideline recommendations by better contextualizing care[18], but routinely adding adjunctive guidance also raises the risk of unsupported extrapolation when outputs are not well calibrated. Consequently, model selection and deployment should prioritize confidence calibration and uncertainty-aware interfaces, employ domain-constrained prompting and alignment techniques, and require human oversight; prospective, workflow-embedded validation is needed to confirm that supplementary outputs improve safety and clinical outcomes. Balancing such supplementary insights against the risks of overextension remains a delicate task that future LLM design must address through domain-constrained prompting and model alignment techniques.

In contrast, readability emerged as a consistent differentiator. DeepSeek-R1 demonstrated significantly lower grade levels and higher Flesch Reading Ease scores compared with Gemini-2.5 Pro, ChatGPT-5 Thinking, and Grok-4 Expert. This suggests that DeepSeek-R1's linguistic structure and lexical distribution may be more optimized for user comprehension—a finding consistent with prior evidence that readability strongly influences the accessibility and usability of medical information for both clinicians and patients.[19,20] We emphasize that these readability metrics were included to evaluate patient-facing suitability rather than to imply superior clinician utility. Clinician-oriented recommendations often require technical terminology, caveats, and concise references that raise measured grade levels while increasing clinical usefulness; therefore, lower grade level is not necessarily preferable for clinician decision support. Therefore, lower grade level is not necessarily

preferable for clinician decision support. Because we did not perform separate clinician- vs. patient-targeted content analyses in this study, the observed readability advantage should be interpreted cautiously: it indicates greater lay readability but does not by itself demonstrate superior clinical decision-making utility.[21,22]

Overall, the findings reinforce that accuracy parity among models does not equate to functional equivalence in clinical utility. Readability, explainability, and user trust are emerging as parallel axes of evaluation that warrant equal consideration when assessing LLMs for healthcare deployment.

This study has several methodological limitations. First, the use of standardized, single-turn scenarios cannot capture the iterative complexity of real clinical practice[23], precluding evaluation of dynamic reasoning and dialogue management. Second, although based on blinded assessment, the small panel of three raters leaves room for rater bias, limiting the reliability and generalizability of the results. Finally, automated readability metrics were used as a proxy for comprehension without direct testing among clinicians or patients, overlooking critical aspects of practical understanding and usability. Future work should adopt longitudinal, interactive designs, engage larger and more diverse rater panels, and employ empirical testing with end-users to properly assess the real-world utility of AI-generated clinical communication.

In conclusion, in this controlled, multidimensional comparison of four state-of-the-art large language models (LLMs) on migraine prophylaxis scenarios derived from the International Headache Society (IHS) criteria, all models demonstrated comparable guideline concordance. However, meaningful disparities emerged in supplementary content and readability; DeepSeek-R1 provided the most comprehensive supplementary information and the most accessible language. Future research should prioritize real-world clinical validation, human-AI collaboration, and domain-specific fine-tuning to ensure these models can safely augment evidence-based medical decision-making.

ACKNOWLEDGEMENTS

We used AI tools—DeepSeek, Gemini, ChatGPT, and Grok—to answer the survey questions, and the full answers are provided in the supplementary materials.

Financial support: This study was supported by the Construction Fund of Key Medical Disciplines of Hangzhou (Grant No. 0020200484), Medical and Health Science Program of Zhejiang Province (Grant: 2025HY0635) and Research on Chronic Disease Management of the National Health Commission (Grant: GWJJMB202510041115).

Conflict of interests: None

REFERENCES

1. Thirunavukarasu AJ, Ting D, Elangovan K, Gutierrez L, Tan TF, Ting D. Large language models in medicine. *Nat Med* 2023; 29(8): 1930-40. DOI: 10.1038/s41591-023-02448-8
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388(13): 1233-9. DOI: 10.1056/NEJMsr2214184
3. Ma B, Yang J, Wong F, *et al.* Artificial intelligence in elderly healthcare: A scoping review. *Ageing Res Rev* 2023; 83: 101808. DOI: 10.1016/j.arr.2022.101808
4. Steiner TJ, Stovner LJ, Jensen R, *et al*. Migraine remains second among the world's causes of disability, and first among young women: findings from GBD2019. *J Headache Pain* 2020; 21(1): 137. DOI: 10.1186/s10194-020-01208-0
5. GBD 2019 Diseases and Injuries Collaborators. 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* 2020; 396(10258): 1204-22. DOI: 10.1016/S0140-6736(20)30925-9
6. Puledda F, Sacco S, Diener HC, *et al.* International Headache Society global practice recommendations for preventive pharmacological treatment of migraine. *Cephalalgia* 2024; 44(9): 3331024241269735. DOI: 10.1177/03331024241269735
7. Fleurence RL, Bian J, Wang X, *et al.* Generative artificial intelligence for health technology assessment: Opportunities, challenges, and policy considerations: An ISPOR Working Group report. *Value Health* 2025; 28(2): 175-83. DOI: 10.1016/j.jval.2024.10.3846
8. McCoy LG, Ci Ng FY, Sauer CM, *et al*. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review. *BMC Med Educ* 2024; 24(1): 1096. DOI: 10.1186/s12909-024-06048-z
9. Tam T, Sivarajkumar S, Kapoor S, *et al.* A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med* 2024; 7(1): 258. DOI: 10.1038/s41746-024-01258-7
10. Hager P, Jungmann F, Holland R, *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024; 30(9): 2613-22. DOI: 10.1038/s41591-024-03097-1
11. Yu E, Chu X, Zhang W, *et al.* Large language models in medicine: Applications, challenges, and future directions. *Int J Med Sci* 2025; 22(11): 2792-801.

DOI: 10.7150/ijms.111780
12. Bianchi O, Willey M, Alvarado CX, *et al*. CARDBiomedBench: A benchmark for evaluating large language model performance in biomedical research: A novel question-and-answer benchmark designed to assess large language models' comprehension of biomedical research, piloted on neurodegenerative diseases. *bioRxiv* 2025; 2025.01.15.633272. DOI: 10.1101/2025.01.15.633272
13. Deng J, Qiu X, Dong C, *et al*. Evaluating ChatGPT and DeepSeek in postdural puncture headache management: a comparative study with international consensus guidelines. *BMC Neurol* 2025; 25(1): 264. DOI: 10.1186/s12883-025-04280-8
14. Maity S, Saikia MJ. Large language models in healthcare and medical applications: A review. *Bioengineering* (Basel) 2025; 12(6): 631. DOI: 10.3390/bioengineering12060631
15. Zhou S, Xu Z, Zhang M, *et al.* Large language models for disease diagnosis: a scoping review. *NPJ Artif Intell* 2025; 1(1): 9. DOI: 10.1038/s44387-025-00011-z
16. Dong C, Qiu X, Deng J, *et al.* Comparative evaluation of large language models in delivering guideline-compliant recommendations for topical NSAID use in musculoskeletal pain: a multidimensional analysis. *Clin Rheumatol* 2025; 44(11): 4703-10. DOI: 10.1007/s10067-025-07640-4
17. Wu HX, Deng JJ, Qiu X, *et al.* Comparative assessment of large language models in diabetic foot infection management: Alignment with IWGDF/IDSA Guidelines. *Front Endocrinol* (Lausanne) 2026:17:1667159. doi: 10.3389/fendo.2026.1667159.
18. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025; 25(1): 117. DOI: 10.1186/s12911-025-02954-4
19. Rooney MK, Santiago G, Perni S, *et al.* Readability of patient education materials from high-impact medical journals: A 20-year analysis. *J Patient Exp* 2021; 8: 2374373521998847. DOI: 10.1177/2374373521998847
20. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T. Readability of written information for patients across 30 years: A systematic review of systematic reviews. *Patient Educ Couns* 2025; 135: 108656. DOI: 10.1016/j.pec.2025.108656
21. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Serv Res* 2022; 22(1): 1148. DOI: 10.1186/s12913-022-08527-9
22. Garcia B, Chua E, Brah HS. The problem of atypicality in LLM-powered psychiatry. *J Med Ethics* 2025; jme-2025-110972. DOI: 10.1136/jme-2025-110972
23. Gallo RJ, Baiocchi M, Savage TR, Chen JH. Establishing best practices in large language model research: an application to repeat prompting. *J Am Med Inform Assoc* 2025; 32(2): 386-90. DOI: 10.1093/jamia/ocae294

Supplementary Table S1. Inter-rater reliability was quantified using ICC[2,1] separately for each model and overall across all four models.

Per-model ICC[2,1]:

Dimension	DeepSeek R1	Gemini 2.5pro	Chat-GPT Think mini	Grok4 Expert
Accuracy	0.676 (95% CI: 0.402–0.796, p<0.001)	0.620 (95% CI: 0.312–0.743, p<0.001)	0.692 (95% CI: 0.391–0.820, p<0.001)	0.557 (95% CI: 0.291–0.691, p<0.001)
Overconclusiveness	0.815 (95% CI: 0.612–0.899, p<0.001)	0.652 (95% CI: 0.286–0.784, p<0.001)	0.577 (95% CI: 0.290–0.698, p<0.001)	0.789 (95% CI: 0.578–0.891, p<0.001)
Supplementary Value	0.616 (95% CI: 0.408–0.713, p<0.001)	0.473 (95% CI: 0.106–0.694, p<0.001)	0.630 (95% CI: 0.323–0.760, p<0.001)	0.582 (95% CI: 0.286–0.697, p<0.001)
Incompleteness	0.699 (95% CI: 0.364–0.822, p<0.001)	0.680 (95% CI: 0.400–0.795, p<0.001)	0.732 (95% CI: 0.354–0.868, p<0.001)	0.513 (95% CI: 0.220–0.673, p<0.001)

Overall ICC[2,1] (all models pooled):

Dimension	Overall ICC[2,1]
Accuracy	0.626 (95% CI: 0.515–0.703, p<0.001)
Overconclusiveness	0.715 (95% CI: 0.613–0.784, p<0.001)
Supplementary Value	0.631 (95% CI: 0.511–0.711, p<0.001)
Incompleteness	0.674 (95% CI: 0.542–0.757, p<0.001)