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.

The AI Response to the 2025ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes: A Comparative Evaluation of Five Large Language Models

Gülşen Genç Tapar

drggenc@gmail.com

Tokat Gaziosmanpaşa University

Ahmet Tuğrul Yılmaz

Tokat Gaziosmanpaşa Üniversitesi

Sefa Erdi ÖMÜR

Tokat Gaziosmanpaşa University

Metin KARAYAKALI

Tokat Gaziosmanpaşa University

Research Article

Keywords: Large language models, artificial intelligence, chronic coronary syndrome, clinical decision support, guideline adherence, Gemini, DeepSeek, Claude, Copilot, ChatGPT, cardiology

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11125999/v1

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

Additional Declarations: No competing interests reported.

The AI Response to the 2025ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes: A Comparative Evaluation of Five Large Language Models

Gülşen GENÇ TAPAR[1], Ahmet TUĞRUL ŞAHİN [2] , Sefa Erdi ÖMÜR[1], Metin KARAYAKALI[1]

1 Department of Cardiology, Tokat Gaziosmanpaşa University, Tokat, Turkey

2 Department of Anesthesiology and Reanimation , Tokat Gaziosmanpaşa University, Tokat, Turkey

Correspondance Author: Gülşen GENÇ TAPAR; drggenc@gmail.com

Abstract

Background and Aim: Cardiovascular diseases, particularly chronic coronary syndromes (CCS), remain a leading cause of morbidity and mortality. Large language models (LLMs) have emerged as promising tools for processing medical information and supporting clinical decision-making. However, systematic comparative evaluation of their performance on current cardiology guidelines remains limited. This study aimed to comparatively evaluate the quality, guideline adherence, and clinical applicability of responses generated by five LLMs to clinical questions derived from the 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes.

Methods: This was a cross-sectional, observational, methodological simulation study. A pool of 20 open-ended clinical questions was developed based on 2025 ACC/AHA/ACEP /NAEMSP /SCAI Guideline for the Managementof Patients With Acute Coronary Syndromes, covering diagnostic and therapeutic domains, and validated by a panel of three independent cardiologists with over 15 years of clinical experience. Questions were submitted to five LLMs Gemini, DeepSeek, Claude, Copilot, and ChatGPT using identical Turkish-language prompts via publicly available web interfaces in April 2026. Model identifiers were removed to ensure evaluator blinding. Five independent expert cardiologists scored each response on five domains Comprehensiveness, Guideline Appropriateness, Understandability, Clinical Applicability, and Overall Scoreusing a 5-point Likert scale. Interrater reliability was assessed via intraclass correlation coefficients (ICC) and Cronbach's alpha. Group comparisons were performed using linear mixed-effects models with Bonferroni correction.

Results: Interrater reliability was acceptable to good across all domains (ICC range: 0.632–0.762; Cronbach's alpha > 0.80 for all domains). The LLM variable exerted a statistically significant effect across all five evaluation domains (all $p < 0.001$). DeepSeek achieved the highest estimated marginal means in all domains, including Overall Score (EMM = 4.87 ± 0.22), significantly outperforming all other models ($p < 0.05$ for all pairwise comparisons). Gemini demonstrated the lowest performance across all domains (Overall Score EMM = 2.75 ± 0.24). ChatGPT and Claude showed comparable performance in all domains (Overall Score: EMM = 3.79 ± 0.22 vs. 3.61 ± 0.23; $p = 1.000$). Diagnostic questions yielded significantly higher Understandability scores than treatment questions (EMM 3.736 vs. 3.512; mean difference = 0.224, 95% CI: 0.067–0.381; $p = 0.005$), while no significant model × question category interaction was observed in any domain.

Conclusion: Although the DeepSeek model demonstrated superior performance specifically within the context of the 2025 AHA Guidelines, this could be attributed to Chain of Thought reasoning. Nevertheless, the overall reliability of large language models in clinical applications remains variable, and future studies should focus not merely on information retrieval, but on their impact on clinical patient management.

Keywords: Large language models; artificial intelligence; chronic coronary syndrome; clinical decision support; guideline adherence; Gemini; DeepSeek; Claude; Copilot; ChatGPT ; cardiology

Introduction

Cardiovascular diseases, particularly chronic coronary syndromes, remain a major cause of morbidity and mortality worldwide. Current clinical decision-making relies heavily on physician expertise, guideline-directed therapies, and static risk scores, which often inadequately accommodate individual patient complexity [1]. Machine learning (ML) and artificial intelligence (AI) specifically reinforcement learning (RL) may augment current physician driven approaches and provide enhanced cardiovascular disease prevention and management [2].

Guidelines published by the American Heart Association (AHA) serve as comprehensive documents considered the gold standard for ensuring standardization in diagnosis and treatment; however, the technical complexity and frequent updates of these texts can hinder their consistent and rapid application in clinical practice [3].

In recent years, artificial intelligence Technologies particularly large language models have demonstrated significant potential to transform the processing and summarization of medical data, as well as clinical decision support processes [4]. The years 2024 and 2025, in particular, have marked a period of notable momentum in AI's ability to reason using medical data and its improved capacity to comprehend clinical guidelines [5]. Utilizing machine learning and AI models as tools to augment physician expertise enables the optimization of personalized treatment plans, especially in high-risk or complex clinical scenarios [2]. However, it has become essential to validate the ability of LLMs to interpret clinical guidelines specifically the most recently published AHA guidelines and to conduct comprehensive studies regarding the reliability of these models [3].

While existing literature indicates that LLMs demonstrate promising results in clinical decision support systems, there is a need for a systematic comparison of performance variations and clinical accuracy levels across different models [3]. This study aims to comparatively evaluate the responses of five different large language models to clinical scenarios derived from the 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes, focusing on scope, guideline adherence, comprehensibility, and clinical applicability.

Materıals and Methods

Study Design and Ethical Statement

This study is a cross-sectional, observational, and methodological simulation study conducted to evaluate the quality, guideline adherence, and clinical utility of responses provided by Large Language Models (LLMs) to clinical questions based on current cardiology guidelines. No data were collected from human or animal participants, nor were patient information or personal health data used; only textual responses generated by artificial intelligence models were evaluated. Therefore, ethical committee approval was not required for this study. Clinical trial number: not applicable.

Question Development Process

The question pool used in the study was developed based on the 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes. A total of 20 open-ended clinical questions were prepared to reflect daily clinical practice.

Three independent cardiologists, each with over 15 years of clinical experience in the field, participated in the processes of developing the questions, ensuring content validity, and verifying alignment with the guidelines. The expert panel evaluated all questions regarding scope, currency, and clinical accuracy, and finalized them through consensus.

The prepared questions were classified into two main categories based on their content:

- **Diagnosis-oriented questions:** Questions covering clinical symptoms, risk assessment, laboratory and imaging findings, and diagnostic algorithms.

- **Treatment-oriented questions:** Questions covering acute and chronic treatment strategies, pharmacological approaches, interventional procedures, and long-term management plans.

Artificial Intelligence Models and Data Collection

All prepared questions were submitted to the following large language models:

- *Gemini (Google LLC, Mountain View, CA, USA)*

- *DeepSeek (DeepSeek Artificial Intelligence Co., Ltd., Hangzhou, China)*

- *Claude (Anthropic PBC, San Francisco, CA, USA)*

- *Copilot (Microsoft Corporation, Redmond, WA, USA)*

- *ChatGPT (OpenAI, San Francisco, CA, USA)*

All queries were conducted in May 2026 using the publicly available, free web versions of the models. To minimize potential contextual learning effects, a new session was initiated for each model, and browser history, cookies, and session data were cleared prior to the queries. The same Turkish prompt was used to query all models, and the resulting responses were recorded.

Blinding and Clinical Evaluation Process

Responses generated by artificial intelligence models were converted into a standardized format, and model names and identifying characteristics were removed to ensure evaluator blinding. Subsequently, the responses were sent via email to five independent expert cardiologists who had not participated in the question generation process and possessed over 15 years of academic and clinical experience in cardiology. The evaluators were given a one-month period to complete the scoring process.

Each response was evaluated using a 5-point ordinal Likert scale across the following five assessment areas:

1. **Comprehensiveness (relative to the guideline):** The extent to which the response includes the key information from the guideline.

2. **Guideline Appropriateness:** The alignment of the presented information with current guideline recommendations.

3. **Understandability:** Language quality, clarity, fluency, and terminological accuracy.

4. **Clinical Applicability:** The usability of the response in actual clinical practice.

5. **Overall Score:** A holistic assessment of the response's general quality.

Higher scores in all areas indicate better performance.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics for Windows, Version 27.0 (IBM Corp., Armonk, NY, USA). A two-sided p-value of <0.05 was considered the threshold for statistical significance.

Inter-rater Reliability Analysis

Intraclass Correlation Coefficient (ICC) analysis was conducted to determine the agreement and reliability among expert raters. The dataset was converted into a wide format, and separate analyses were performed for each assessment area. The ICC calculations employed a two-way mixed-effects model, an absolute agreement definition, and an average measures approach. Results are reported alongside 95% confidence intervals.

Additionally, Cronbach's Alpha coefficients were calculated for each assessment area to evaluate the internal consistency of the rater panel.

Comparison of Artificial Intelligence Model Performance

Linear Mixed-Effects Models were used to compare the performance of the artificial intelligence models. A separate model was constructed for each evaluation area.

In the models, the following were defined as fixed effects:

- Artificial intelligence model (Gemini, DeepSeek, Claude, Copilot, and ChatGPT),
- Question category (Diagnosis/Treatment),
- Artificial intelligence model × question category interaction.

Scoring discrepancies among evaluating physicians and the dependency structure arising from repeated evaluations of the same question were incorporated into the model as random effects. This approach aimed to reveal the true performance differences of the artificial intelligence models by controlling for the individual scoring tendencies of the evaluators and question-based variability.

Model parameters were estimated using the Restricted Maximum Likelihood (REML) method, and degrees of freedom were calculated using the Satterthwaite approximation. The significance of fixed effects was assessed using Type III F-tests. The Bonferroni correction was applied for multiple pairwise comparisons. Results are reported alongside Estimated Marginal Means (EMM) and standard error (SE) values.

Results

Findings on Inter-rater Agreement and Reliability

The agreement among five expert cardiologists in scoring the responses provided by artificial intelligence models was evaluated using intraclass correlation coefficient (ICC) analysis based on a two-way mixed-effects model, absolute agreement, and average measures approach. ICC values were found to be statistically significant across all evaluation domains ($p<0.001$ for all analyses) (Table 1).

According to the ICC analysis, the highest level of inter-rater agreement was observed in the domains of Scope (ICC=0.762; 95% CI: 0.607–0.851) and Clinical Applicability (ICC=0.757; 95% CI: 0.583–0.852). Acceptable levels of inter-rater agreement were also found for the Overall Score (ICC=0.733; 95% CI: 0.544–0.836), Clarity (ICC=0.730; 95% CI: 0.542–0.834), and Guideline Adherence (ICC=0.632; 95% CI: 0.324–0.788). Overall, the ICC values indicate a level of consensus among expert raters ranging from moderate to good.

In the internal consistency analysis, Cronbach's alpha coefficients were found to be 0.836 for Scope, 0.816 for Guideline Adherence, 0.824 for Clarity, 0.842 for Clinical Applicability, and 0.827 for the Total Score. The fact that Cronbach's alpha values exceeded 0.80 across all assessment domains indicates that the expert panel's scoring system possesses a high level of internal consistency and that the resulting assessments are reliable.

Evaluation of Artificial Intelligence Model Performance Using Linear Mixed-Effects Models

Responses provided by five different large language models (LLMs) to questions derived from the 2023 ESC Guidelines for the management of acute coronary syndromes guidelines were analyzed using linear mixed-effects models, accounting for scoring tendencies among evaluating physicians and question-based variability. In the model, evaluating physicians were defined as a random effect (random intercept) to control for systematic scoring differences between evaluators. Consequently, the results were adjusted to reflect the true performance differences of the artificial intelligence models, independent of individual evaluator biases. Type III fixed-effect tests demonstrated that the artificial intelligence model variable had a statistically significant effect across all evaluation domains, including Comprehensiveness ($F=134.10$; $p<0.001$), Guideline Adherence ($F=96.36$; $p<0.001$), Understandability ($F=93.42$; $p<0.001$), Clinical Applicability ($F=107.29$; $p<0.001$), and Overall Score ($F=114.13$; $p<0.001$). Bonferroni-corrected pairwise comparisons revealed distinct performance differences among the models (Table 2).

Regarding the scope dimension, the DeepSeek model achieved the highest adjusted mean score (EMM=4.89±0.17) and significantly outperformed all other models ($p<0.001$ for all comparisons). The lowest score was observed with the Gemini model (EMM=2.72±0.20). Within the intermediate performance group, no significant difference was found between ChatGPT (EMM=3.74±0.17) and Claude (EMM=3.53±0.18) ($p=0.694$). Similarly, no significant difference was observed between Claude and Copilot (EMM=3.25±0.18) ($p=0.237$). In contrast, the ChatGPT model scored significantly higher than the Copilot model ($p<0.001$). These findings indicate that the Claude model performed statistically similarly to both ChatGPT and Copilot, whereas ChatGPT and Copilot differed from each other.

A similar distribution was observed in the areas of guideline adherence and clinical applicability. While the DeepSeek model achieved the highest scores in both assessment areas (EMM=4.79±0.35 and EMM=4.83±0.23, respectively), the Gemini model received the lowest scores (EMM=2.84±0.36 and EMM=2.56±0.25, respectively). The Claude model demonstrated statistically similar performance to both ChatGPT and Copilot across both dimensions, whereas the difference between ChatGPT and Copilot remained significant even after Bonferroni correction ($p<0.05$).

The DeepSeek model received the highest score in terms of understandability (EMM=4.87±0.25). No significant difference was found between ChatGPT (EMM=3.74±0.25) and Claude (EMM=3.60±0.26) ($p=1.000$). While both models scored higher than Gemini, the Copilot model (EMM=3.16±0.26) demonstrated intermediate performance. The difference between Copilot and Gemini was not found to be statistically significant ($p=0.061$).

When overall scores were evaluated, the DeepSeek model demonstrated the highest performance (EMM=4.87±0.22) and scored significantly higher than all other models ($p<0.05$ for all comparisons). While no significant difference was found between ChatGPT (EMM=3.79±0.22) and Claude (EMM=3.61±0.23) ($p=1.000$), both models scored higher than Copilot (EMM=3.25±0.23). The Gemini model exhibited the lowest overall performance, with an EMM of 2.75±0.24.

Analysis of the Effect of Question Category (Diagnosis/Treatment) and Interactions

Diagnosis- and treatment-oriented question categories were included in the model as fixed effects to evaluate the models' performance across different types of clinical content. The effect of the question category on Comprehensiveness (p=0.168), Guideline Adherence (p=0.121), Clinical Applicability (p=0.051), and Overall Score (p=0.136) was not found to be statistically significant. This result indicates that the models' overall performance levels are largely independent of whether the question is diagnosis- or treatment-oriented. Although the mean scores for diagnostic questions were higher than those for treatment questions across all evaluation domains, this difference reached statistical significance only in the understandability dimension. A significant effect of question category was observed regarding understandability (F=7.821; p=0.005). Understandability scores for diagnosis-oriented questions (EMM=3.736±0.250) were found to be significantly higher than those for treatment-oriented questions (EMM=3.512±0.244). The mean difference was 0.224 points (95% CI: 0.067–0.381), indicating that understandability scores for diagnosis-oriented questions were higher than those for treatment oriented questions.

The interaction between the artificial intelligence model and the question category was found to be non-significant across all evaluation domains: Scope (p=0.991), Adherence to Guidelines (p=0.842), Comprehensibility (p=0.540), Clinical Applicability (p=0.338), and Overall Score (p=0.635). This finding indicates that the relative performance ranking of the models remained consistent across diagnostic and treatment questions and that the question category did not significantly alter the differences in model performance.

One of the study's key methodological strengths is that inter-rater scoring differences were controlled for as a random effect within the linear mixed model. This approach reduced the likelihood that performance differences between models stemmed from the scoring tendencies of a specific evaluator.

Discussion

This study comparatively analyzed the clinical decision-support capabilities of five different large language models in light of the 2025 AHA Guidelines for the management of acute coronary syndromes. The findings indicate that the DeepSeek model demonstrated statistically significant superiority across all evaluation domains (Overall Score EMM = 4.87 ± 0.22). This performance aligns with current data in the literature; the DeepSeek-R1 model's "Chain of Thought" reasoning capability, "Mixture of Experts" architecture, and integration of reinforcement learning enable it to achieve higher accuracy rates than its competitors in complex medical scenarios [6].Furthermore, compared to traditional dense models, models utilizing the MoE architecture have been reported to exhibit superior reasoning efficiency and performance, independent of the parameter count [7].This explains why, in our study, DeepSeek was able to simulate the multi-step logical processes required by clinical guidelines more successfully than the other models.

Another critical finding of our study is that diagnosis-oriented questions received higher scores for comprehensibility compared to treatment-oriented questions. This reflects a distinction within the literature on clinical decision support systems: while diagnostic processes involve logical reasoning such as differential diagnosis formulation and Bayesian inference that models encounter more frequently in training data [3], treatment processes rely on rigid protocols structured as "decision trees" within clinical guidelines [9]. Existing

literature highlights that many LLMs struggle to adhere to these structured treatment protocols and fail to integrate effectively into clinical workflows [10]. Our findings confirm that DeepSeek demonstrates greater robustness and consistency than other models in following these structured protocols, yet they also corroborate the general trend of performance degradation across the LLM ecosystem when transitioning from diagnosis to treatment.

The low performance of the Gemini model observed in our study highlights the challenges general-purpose models face in adapting to specific medical guidelines [11]. The literature indicates that LLMs have limited ability to adapt to precise, region or guideline-specific cut-off values, which may pose clinical safety risks [10]. Consequently, DeepSeek's success demonstrates that medical reasoning capability is more critical than the parameter volume of a general purpose model [7].

Limitations

Several limitations of this study must be acknowledged. First, the questions in the study were formulated in Turkish; while this reflects global patient information-seeking behavior, variations in model performance may occur due to differences in linguistic optimization. Second, although inter-rater reliability was demonstrated using the intraclass correlation coefficient (ICC) analysis, the evaluation process relied on expert opinion, which may not entirely eliminate subjective bias. Third, the models used represent specific publicly available versions from a particular timeframe; given the rapid evolution of language models, these findings should be interpreted with regard to the time at which the study was conducted.

Conclusion

In conclusion, although the DeepSeek model demonstrated superior performance regarding the 2025 AHA guidelines, the overall reliability of large language models in clinical practice remains variable. Adherence to guidelines particularly regarding the following of treatment protocolsis a fundamental requirement for the safe clinical use of artificial intelligence models. Future studies should focus on the impact of these models on clinical patient management, rather than solely on information presentation.

References

1.Kumar, A., M. Gupta, A. Varshney, S. Kumari and R. N. Sahu (2024). "Artificial Intelligence in Cardiac Healthcare: Advancements in Managing Coronary Artery Disease and Acute Coronary Syndrome." Disease and Health Research: New Insights Vol. 10: 91-115.

2.Biondi-Zoccai, G., A. Mahajan, D. Powell, M. Peruzzi, R. Carnevale and G. Frati (2025). "Advancing cardiovascular care through actionable AI innovation." NPJ Digital Medicine **8**(1): 249.

3. Quer, G. and E. J. Topol (2024). "The potential for large language models to transform cardiovascular medicine." The Lancet Digital Health **6**(10): e767-e771.

4. Gendler, M., G. N. Nadkarni, K. Sudri, M. Cohen-Shelly, B. S. Glicksberg, O. Efros, S. Soffer and E. Klang (2026). "Large Language Models in Cardiology: Systematic Review." JMIR cardio **10**(1): e76734.

5. Khattak, S., J. Townend, N. Khan and M. Thomas (2026). "Rapid improvement in ability of AI to reason using clinical guidelines." European Heart Journal-Digital Health **7**(Supplement_1): ztaf143. 056.

6. Hoyt, R., D. Knight, M. Haider and M. Bajwa (2025). "Evaluating a Large Reasoning Model's Performance on Open-Ended Medical Scenarios." medRxiv: 2025.2004. 2029.25326666.

7. Tseng, C.-Y., D. Zhang, T. Wang, H. Luo, L. Chen, J. Huang, J. Guan, J. Hao, J. Song and X. Song (2025). "47b mixture-of-experts beats 671b dense models on chinese medical examinations." arXiv preprint arXiv:2511.21701.

8. Savage, T., A. Nayak, R. Gallo, E. Rangan and J. H. Chen (2024). "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine." NPJ Digital Medicine **7**(1): 20.

9. Li, X., M. Gao, Y. Hao, T. Li, G. Wan, Z. Wang and Y. Wang (2025). "MedGUIDE: benchmarking clinical decision-making in large language models." arXiv preprint arXiv:2505.11613.

10. Hager, P., F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren and G. Kaissis (2024). "Evaluation and mitigation of the limitations of large language models in clinical decision-making." Nature medicine **30**(9): 2613-2622.

11. Park, Y.-J., A. Pillai, J. Deng, E. Guo, M. Gupta, M. Paget and C. Naugler (2024). "Assessing the research landscape and clinical utility of large language models: a scoping review." BMC Medical Informatics and Decision Making **24**(1): 72.

Supplementary Materials

Table 1. Inter-Rater Reliability and Agreement Analysis Among Cardiologist Evaluators; Table 2. Comparative Performance of Large Language Models Across Evaluation Domains Based on Linear Mixed-Effects Model Analysis.

Author Contributions

Conceptualization, G.G.T. and S.E.Ö.; methodology, G.G.T., S.E.Ö. validation, M.K. and S.E.Ö.; formal analysis, A.T.Ş.; investigation, G.G.T., M.K.; data curation, M.K. ; writing -original draft preparation, G.G.T.; writing- review and editing, all authors; visualization, G.G.T; supervision, G.G.T.; project administration, G.G.T.,S.E.Ö. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived for this study because it was conducted as an in silico analysis using publicly available artificial intelligence models and did not involve human participants, patient data, or biological material. All data were analyzed and reported in accordance with principles of scientific integrity and publication ethics. The evaluated AI models were not used as clinical decision-support tools and were assessed exclusively for research purposes.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

During the preparation of this manuscript, the authors used ChatGPT (OpenAI, San Francisco, CA, USA; **https://chat.openai.com**) for language editing and figure visualization based on the authors' original data. The authors reviewed and edited all outputs and take full responsibility for the content.

Conflicts of Interest

The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Table 1. Inter-Rater Reliability and Agreement Analysis Among Cardiologist Evaluators

Evaluation Domain	Cronbach's Alpha	ICC (Average Measures)	95% Confidence Interval	F Value	p Value
Comprehensiveness	0.836	0.762	0.607–0.851	6.112	<0.001
Guideline Appropriateness	0.816	0.632	0.324–0.788	5.428	<0.001
Understandability	0.824	0.730	0.542–0.834	5.695	<0.001
Clinical Applicability	0.842	0.757	0.583–0.852	6.339	<0.001
Overall Score	0.827	0.733	0.544–0.836	5.787	<0.001

ICC: Intraclass Correlation Coefficient. ICC values were calculated using a two-way mixed-effects model with an absolute agreement definition based on average measures.

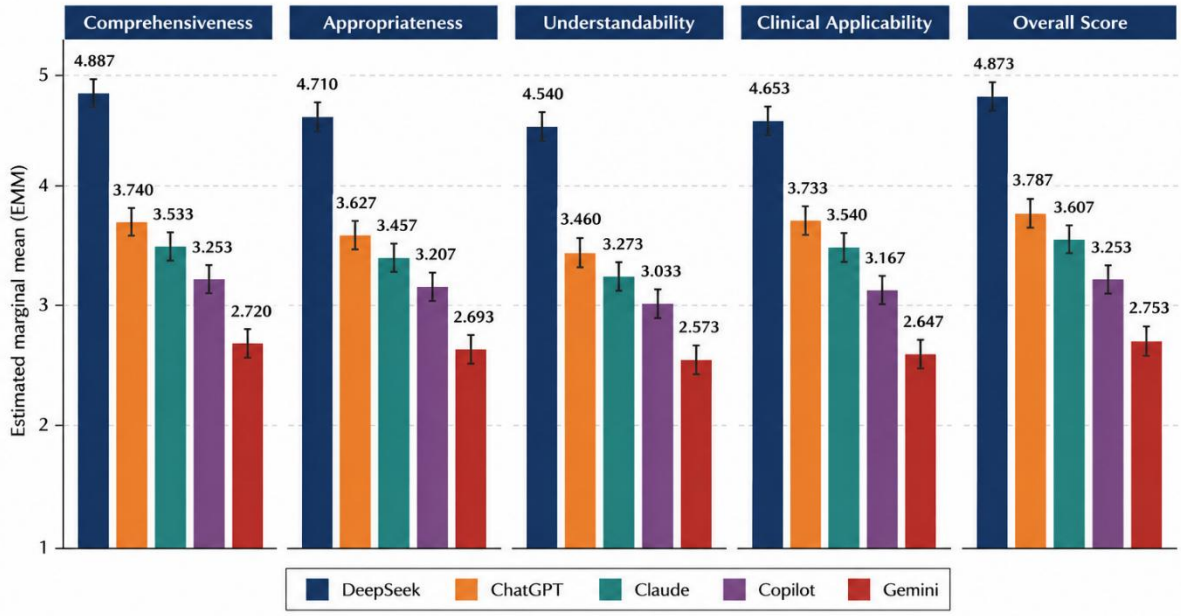
Table 2. Comparative Performance of Large Language Models Across Evaluation Domains Based on Linear Mixed-Effects Model Analysis

Domain	Gemini (EMM ± SE)	DeepSeek (EMM ± SE)	Claude (EMM ± SE)	Copilot (EMM ± SE)	ChatGPT (EMM ± SE)	F value	p value
Comprehensiveness	2.72 ± 0.20^{d}	4.89 ± 0.17^{a}	3.53 ± 0.18^{bc}	3.25 ± 0.18^{c}	3.74 ± 0.17^{b}	134.10	<0.001
Guideline Appropriateness	2.84 ± 0.36^{d}	4.79 ± 0.35^{a}	3.55 ± 0.35^{bc}	3.27 ± 0.36^{c}	3.69 ± 0.35^{b}	96.36	<0.001
Understandability	2.75 ± 0.27^{c}	4.87 ± 0.25^{a}	3.60 ± 0.26^{b}	3.16 ± 0.26^{c}	3.74 ± 0.25^{b}	93.42	<0.001
Clinical Applicability	2.56 ± 0.25^{d}	4.83 ± 0.23^{a}	3.54 ± 0.24^{bc}	3.22 ± 0.24^{c}	3.71 ± 0.23^{b}	107.29	<0.001
Overall Score	2.75 ± 0.24^{d}	4.87 ± 0.22^{a}	3.61 ± 0.23^{b}	3.25 ± 0.23^{c}	3.79 ± 0.22^{b}	114.13	<0.001

EMM: Estimated Marginal Mean; SE: Standard Error.

Values are presented as estimated marginal means ± standard errors derived from linear mixed-effects models. F and p values indicate the overall effect of AI model. Pairwise comparisons were adjusted using the Bonferroni correction. Within each row, values sharing at least one common superscript letter are not significantly different, whereas values with no common superscript letters differ significantly ($p < 0.05$).

Figure 1. Estimated marginal means (± SE) for AI models across evaluation subdomains (Linear mixed model)



Scores are on a 1–5 Likert scale (1 = strongly disagree, 5 = strongly agree).
Error bars represent standard error (SE).
The main effect of AI model was significant for all subdomains (all $p < 0.001$).