Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
51 (2) 239-246, 2025
DOI: https://doi.org/10.32708/uutfd.1718121

ORIGINAL RESEARCH

Performance of Generative AI Models on Cardiology Practice in Emergency Service: A Pilot Evaluation of GPT-4.o and Gemini-1.5-Flash

Şeyda GÜNAY-POLATKAN[1], Deniz SIĞIRLI[2], Vahide Aslihan DURAK[3], Çetin ALAK[1], İrem IRİS KAN[4]

[1] Bursa Uludag University, Faculty of Medicine, Departmant of Cardiology, Bursa, Türkiye.
[2] Bursa Uludag University, Faculty of Medicine, Departmant of Biostatistics, Bursa, Türkiye.
[3] Bursa Uludag University, Faculty of Medicine, Departmant of Emergency Medicine, Bursa, Türkiye.
[4] Bursa Uludag University, Faculty of Medicine, Departmant of Cardiovascular surgery, Bursa, Türkiye.

ABSTRACT

In healthcare, emergent clinical decision-making is complex and large language models (LLMs) may enhance both the quality and efficiency of care by aiding physicians. Case scenario-based multiple choice questions (CS-MCQs) are valuable for testing analytical skills and knowledge integration. Moreover, readability is as important as content accuracy. This study aims to compare the diagnostic and treatment capabilities of GPT-4.o and Gemini-1.5-Flash and to evaluate the readability of the responses for cardiac emergencies. A total of 70 single-answer MCQs were randomly selected from the Medscape Case Challenges and ECG Challenges series. The questions were about cardiac emergencies and were further categorized into four subgroups according to whether the question included a case presentation or an image, or not. ChatGPT and Gemini platforms were used to assess the selected questions. The Flesch–Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) scores were utilized to evaluate the readability of the responses. GPT-4.o had a correct response rate of 65.7%, outperforming Gemini-1.5-Flash, which had a 58.6% correct response rate (p=0.010). When comparing by question type, GPT-4.o was inferior to Gemini-1.5-Flash only for non-case questions (52.5% vs. 62.5%, p=0.011). For all other question types, there were no significant performance differences between the two models (p>0.05). Both models performed better on easy questions compared to difficult ones, and on questions without images compared to those with images. Additionally, while GPT-4.o performed better on case questions than non-case questions. Gemini-1.5-Flash's FRE score was higher than GPT-4.o's (median [min-max], 23.75 [0-64.60] vs. 17.0 [0-56.60], p<0.001). Although on the whole GPT-4.o outperformed Gemini-1.5-Flash, both models demonstrated an ability to comprehend the case scenarios and provided reasonable answers.

Keywords: Cardiology. Decision making. Artificial intelligence. GPT-4.o. Gemini-1.5-Flash.

Kardiyak Acil Durumların Yönetiminde ChatGPT ve Gemini

ÖZET

Sağlık hizmetlerinde, acil klinik karar alma karmaşıktır ve büyük dil modelleri (LLM'ler) hekimlere yardımcı olarak hem bakımın kalitesini hem de verimliliğini artırabilir. Vaka senaryosuna dayalı çoktan seçmeli sorular (VS-ÇSS), analitik becerileri ve bilgi bütünleştirmeyi test etmek için değerlidir. Ayrıca, okunabilirlik, içerik doğruluğu kadar önemlidir. Bu çalışma, GPT-4.o ve Gemini-1.5-Flash'ın tanı ve tedavi yeteneklerini karşılaştırmayı ve kardiyak acil durumlar için yanıtların okunabilirliğini değerlendirmeyi amaçlamaktadır. Medscape Vaka Zorlukları ve EKG Zorlukları serilerinden toplam 70 tek cevaplı ÇSS rastgele seçildi. Sorular kardiyak acil durumlarla ilgiliydi ve sorunun bir vaka sunumu veya bir görüntü içerip içermemesine göre dört alt gruba ayrıldı. Seçilen soruları değerlendirmek için CahtGPT ve Gemini platformları kullanıldı. Yanıtların okunabilirliğini değerlendirmek için Flesch-Kincaid Sınıf Düzeyi (FKGL) ve Flesch Okuma Kolaylığı (FRE) puanları kullanıldı. GPT-4.o'nun doğru yanıt oranı %65,7'ydi ve %58,6 doğru yanıt oranına sahip Gemini-1.5-Flash'ı geride bıraktı (p=0,010). Soru türüne göre karşılaştırıldığında, GPT-4.o yalnızca vaka dışı sorularda Gemini-1.5-Flash'tan daha düşüktü (%52,5'e karşı %62,5, p=0,011). Diğer tüm soru türleri için, iki model arasında önemli bir performans farkı yoktu (p>0,05). Her iki model de kolay sorularda zor sorulara göre ve resimsiz sorularda resimli sorulara göre daha iyi performans gösterdi. Ek olarak, GPT-4.o vaka dışı sorulara göre vaka sorularında daha iyi performans gösterdi. Gemini-1.5-Flash'ın FRE puanı GPT-4.o'dan daha yüksekti (ortanca [min-maks], 23.75 [0-64.60] - 17.0 [0-56.60], p<0.001). Her ne kadar toplamda GPT-4.o, Gemini-1.5-Flash'tan daha iyi performans gösterse de, her iki model de durum senaryolarını anlama becerisi gösterdi ve makul yanıtlar sağladı.

Anahtar Kelimeler: Kardiyoloji. Karar verme. Yapay zeka. GPT-4.o. Gemini-1.5-Flash.

Date Received: 12.June.2025
Date Accepted: 2.July.2025

Dr. Şeyda GÜNAY-POLATKAN
Bursa Uludag University, Faculty of Medicine,
Departmant of Cardiology
Post code: 16059, Bursa, Turkey
e-mail: seydagunay@uludag.edu.tr

AUTHORS' ORCID INFORMATION
Şeyda GÜNAY-POLATKAN: 0000-0003-0012-345X
Deniz SIĞIRLI: 0000-0002-4006-3263
Vahide Aslıhan DURAK: 0000-0003-0836-7862
Çetin ALAK: 0000-0003-1875-2078
İrem İRİS KAN: 0000-0002-1600-9531

The advancement of digital technologies has significantly impacted various aspects of daily life, including how people access health information globally. Generative Pre-trained Transformer (GPT) models, designed for deep learning tasks like text generation, language modeling, and text completion, have become essential in this context[1-3]. The integration of deep learning (DL) with natural language processing (NLP) and the availability of large datasets have led to the emergence of large language models (LLMs)[4]. In healthcare, where clinical decision-making is becoming increasingly complex, LLMs have the potential to enhance both the quality and efficiency of care by aiding physicians[5]. Accurate evaluation in emergency departments (EDs) depends on prompt disease diagnosis and treatment. In order to treat patients whose symptoms closely match those of a certain specialty, emergency physicians (EPs) may need the assistance of an attending physician. The consultation procedure can be a significant burden for the medical community due to the 24 hours a day unavailability of cardiologists and other specialists in emergency departments. In addition to needing help reading a patient's ECG, emergency department doctors may occasionally need to consult in order to diagnose and treat a patient.

AI-powered conversational agents can simulate human-like interactions and are useful for delivering medical information. OpenAI introduced ChatGPT, powered by GPT-3.5, in 2022 as a general-purpose AI chatbot[6,7], followed by GPT-4.0 in March 2023 and the more advanced GPT-4.o, which can handle both image and text inputs, in May 2024[8,9]. Google's Gemini, previously known as Bard, is another generative AI chatbot, with studies comparing the performance of ChatGPT and Gemini across various medical specialties[10-16].

Multiple-choice question (MCQ) exams are widely used in educational assessments across many disciplines, including medicine. These questions can either be stand-alone (questioning knowledge, not including a case presentation; named as non-case questions in this study) or based on patient scenarios, which include laboratory results, vital signs, and diagnostic tests (named as case questions in this study). Case scenario-based MCQs (CS-MCQs) are particularly valuable for testing analytical skills, problem-solving abilities and knowledge integration, making them ideal for problem-based learning (PBL)[17]. When evaluating the performance of AI in answering such questions, readability is also as important as content accuracy. Readability can be assessed using objective, quantitative formulas like the Flesch-Kincaid Grade Level (FKGL) and the Flesch Reading Ease (FRE) score[18].

The purpose of this study was to assess the success of using LLMs rather than cardiology consultations for cardiology cases and/or ECG interpretation comparing the diagnostic and treatment capabilities of GPT-4.o and Gemini-1.5-Flash using cardiology MCQs sourced from Medscape which is one of the leading online global destination for physicians and healthcare professionals worldwide providing quick access to medical information in daily practice. Medscape website includes both stand-alone questions and case presentations with and without an ECG image. Additionally, since the readability and verbosity of the responses are important in daily practice, the responses generated by both models were evaluated using the FKGL and FRE metrics[18].

Material and Method

Study design

Medscape provides comprehensive medical information for healthcare professionals, including cardiology-related content. For this study, all cardiology questions from the Medscape Case Challenges[19] and ECG Challenges[20] series were screened to select cardiac emergent issues. Any questions that contained visual elements other than ECG images, such as radiological or clinical images were excluded. A total of 70 freely accessible single-answer MCQs were randomly selected from these sources[19,20]. Alongside the correct answer, real life data regarding the percentage of human respondents who answered correctly were also available. This data allowed us to classify each question as either difficult (correct response rate below 60%) or easy (correct response rate 60% or higher). Questions that included patient history, physical examination, and laboratory findings were categorized as case questions, while others were labeled non-case questions. The questions were further classified as image or non-image based on the inclusion of an ECG image.

The questions were further categorized into four subgroups: those included neither a case presentation nor an image (group 1), those with an image but no case presentation (group 2), those with a case presentation but without an image (group 3), and those containing both a case presentation and an image (group 4).

Evaluation tools:

Both OpenAI's GPT-4.o and Google's Gemini-1.5-Flash were used to assess the selected questions. Each question was inputted identically into both platforms only once, and their responses were categorized as correct or incorrect based on Medscape's answer key. Additionally, the percentage of human respondents who answered correctly was noted for each question.

The Flesch–Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) scores were utilized to evaluate the readability and verbosity of the responses generated by GPT-4.o and Gemini-1.5-Flash. The FKGL estimates the reading grade level of a text, while the FRE assigns a score between 1 and 100, with higher scores indicating easier readability (Table I). Both metrics account for the number of sentences, words, and syllables in the text to measure verbosity. The FKGL is calculated using the formula: FKGL = (0.39 × [Total Words / Total Sentences] + 11.8 × [Total Syllables / Total Words] − 15.59), and the FRE score is determined using the equation: FRE = (206.835 − 1.015 × [Total Words / Total Sentences] − 84.6 × [Total Syllables / Total Words]). The number of words, syllables, and sentences in each text was automatically calculated using the website https://readability-score.com/18, 21, 22.

Table I. Interpretation of Flesch Reading Ease Score (22)

Flesch Reading Ease Score	Readability Level	Estimated Reading Grade
0-30	very difficult	college graduate
30-50	difficult	college
50-60	fairly difficult	10th to 12th grade
60-70	standard	8th or 9th grader
70-80	fairly easy	7th grader
80-90	easy	6th grader
90-100	very easy	5th grader

Statistical Analyses:

The normality of the data was tested using the Shapiro-Wilk test. Normally distributed data were presented with mean±standard deviation while non-normal data presented with median (minimum-maximum) values. For non-normally distributed variables, the Mann-Whitney U test was employed to compare two independent groups. To compare two depedendent groups paired t test was used for normally disributed data, while the Wilcoxon test was used for non-normal data. Categorical variables were analyzed using Pearson chi-square and Fisher's exact chi-square tests, with the data reported as n(%). A significance level of 0.05 was considered for two-sided hypothesis tests. All statistical analyses were conducted using IBM SPSS Statistics (Version 28.0, IBM Corp, Armonk, NY).

Results

The study included 70 randomly selected questions from the Medscape Case Challenges and ECG Challenges[19,20]. The questions represented various types: some included a case presentation or image, while others did not. Additionally, they were categorized by difficulty, with 36 hard and 34 easy questions, 30 case questions, 40 non-case questions, 31 image questions, and 39 non-image questions. GPT-4.o had a correct response rate of 65.7%, outperforming Gemini-1.5-Flash, which had a 58.6% correct response rate—a significant difference of 7.1% (p=0.010, Table II). When comparing by question type, GPT-4.o was inferior to Gemini-1.5-Flash only for non-case questions (52.5% vs. 62.5%, p=0.011). For all other question types, there were no significant performance differences between the two models (p>0.05, Table III).

Table II. Comparison of GPT-4.o and Gemini-1.5-Flash for all of the questions

	GPT-4.o (n,%)	Gemini-1.5-Flash (n,%)	p value
Correct	46 (65.7)	41 (58.6)	0.010
Incorrect	24 (34.3)	29 (41.4)	

Table III. Distribution of correct response percentage of GPT-4.o and Gemini-1.5-Flash according to questions types

Question type		GPT-4.o n (%)	Gemini-1.5-Flash n (%)	p-value
Difficulty level	Hard	17 (47.2)	16 (44.4)	0.332
	Easy	29 (85.3)	25 (73.5)	0.102
Case presentation	Case	25 (83.3)	16 (53.3)	0.157
	Non-case	21 (52.5)	25 (62.5)	**0.011**
Image presence	Image	11 (35.5)	14 (45.2)	0.153
	Non-image	35 (89.7)	27 (69.2)	0.573

Both models performed better on easy questions compared to difficult ones, and on questions without images compared to those with images. Additionally, while GPT-4.o performed better on case questions than non-case questions, the presence of a case presentation had no impact on Gemini's performance. Detailed comparison results for each question type are displayed in Figures 1 and 2 for GPT-4.o and Gemini-1.5-Flash, respectively.

When the questions were divided into subgroups based on the presence of a case or image, the group 1 (without a case presentation and an image) contained 18 questions, group 2 (with an image, without a case presentation) contained 22 questions, group 3 (with a case presentation, without an image) contained 21 questions and group 4 (with a case presentation and an image) contained 9 questions. GPT-4.o provided 46 correct answers, while Gemini-1.5-Flash provided 41 correct answers. These responses are distributed across the subgroups, as shown in Figure 3. A significant difference was found in GPT-4.o's correct response rate across the subgroups (p<0.001), but no significant difference was found for Gemini-1.5-Flash (p=0.076, Table IV). Pairwise comparisons revealed that groups 1 vs. 2, 2 vs. 3, and 2 vs. 4 had significant differences (p<0.001, p<0.001, and p=0.038, respectively). Group 2 (with an image, without a case presentation) showed the lowest performance from GPT-4.o.

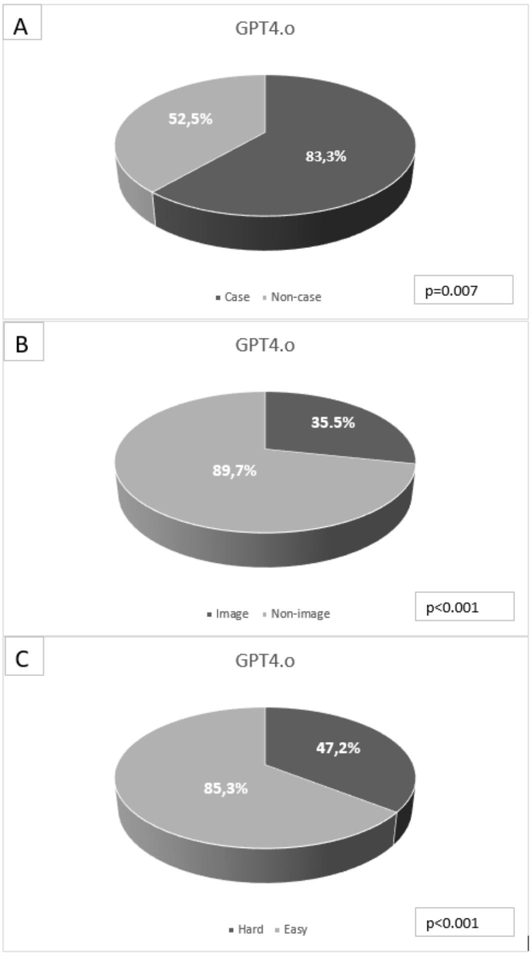

Figure 1:

Comparisons of correct response percentages of GPT-4.o for each question type

Table IV. Comparison of GPT-4.o and Gemini-1.5-Flash among question subgroups regarding correct response percentage

		GPT-4.o n (%)	**Gemini-1.5-Flash** n (%)
Question subgroup	Group 1	16 (34.8)	15 (36.6)
	Group 2	5 (10.9%)	10 (24.4%)
	Group 3	19 (41.3%)	12 (29.3%)
	Group 4	6 (13.0%)	4 (9.8%)
p-value		<0.001	0.076

Group 1; without a case presentation and an image, Group 2; with an image, without a case presentation, Group 3; with a case presentaion, without an image, Group 4; with an image and a case presentation

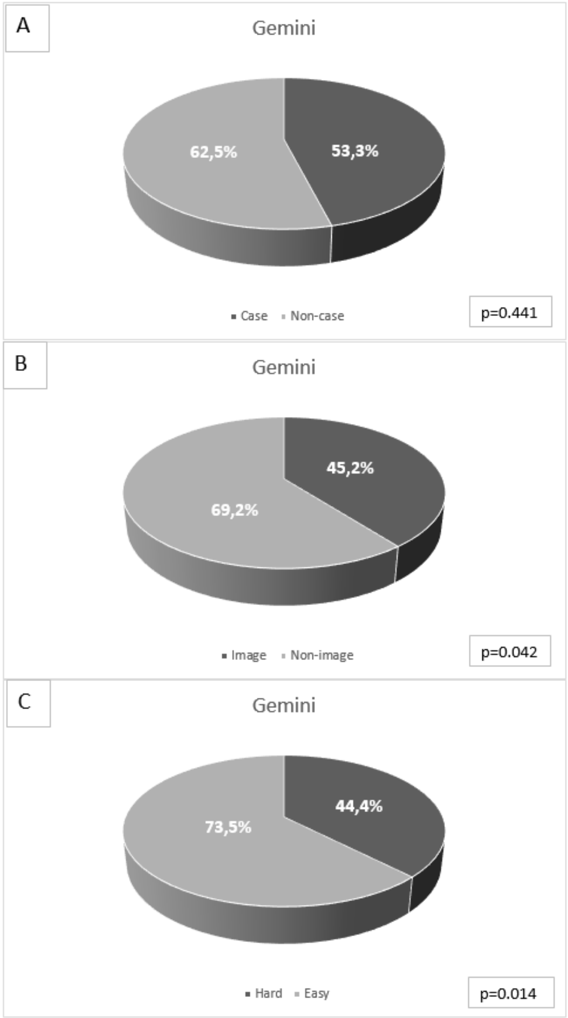

Figure 2:

Comparisons of correct response rates of Gemini-1.5-Flash for each question type

None of the questions answered correctly by GPT-4.o or Gemini-1.5-Flash were answered correctly by all human participants. However, some human respondents answered questions correctly that both AI models answered incorrectly. The median (minimum-maximum) correct response rate for human participants across the 70 questions was 58.50% (15.00-94.00). Human participants performed better on questions where GPT-4.o gave the correct answer (median [min-max], 63.50% [15.00-94.00]) compared to those where GPT-4.o answered incorrectly (median [min-max], 47.50% [17.00-74.00], p<0.001). A similar difference was found for Gemini-1.5-Flash's correct and incorrect answers in terms of human performance (p=0.001, Table V).

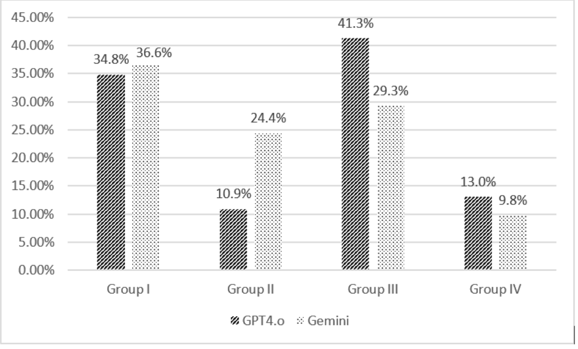

Figure 3:
The distribution of correct responses given by GPT-4.o (n=46) and Gemini-1.5-Flash (n=41) into subgroups (Group 1; without a case presentation and an image, Group 2; with an image, without a case presentation, Group 3; with a case presentaion, without an image, Group 4; with an image and a case presentation)

Table V. Comparison of the correct response percentage of human participants in questions answered correctly and incorrectly by GPT 4.o and Gemini-1.5-Flash

	correct	incorrect	p- value
GPT-4.o	63.5 (15-94)	47.5 (17-74)	<0.001
Gemini-1.5-Flash	63 (38-94)	47 (15-92)	0.001

Data presented as median (minimum-maximum)

Regarding readability and verbosity, GPT-4.o had a higher FKGL than Gemini-1.5-Flash (mean ± SD, 14.79 ± 2.76 vs. 13.87 ± 3.56, p=0.023). Both models produced responses with a FRE score below 30, indicating that their texts were at the college graduate level. However, Gemini-1.5-Flash’s FRE score was higher than GPT-4.o’s (median [min-max], 23.75 [0-64.60] vs. 17.0 [0-56.60], p<0.001, Table VI).

Table VI. Comparison of readability and verbosity of responses given by GPT-4.o and Gemini-1.5-Flash

	GPT-4.o	Gemini-1.5-Flash	p value
FKGL	14.79 ± 2.76	13.87 ±3.56	0.023
FRE score*	17.00 (0-56.60)	23.75 (0-64.60)	<0.001

*Data presented as median (minimum-maximum),
FKGL: Flesch–Kincaid Grade Level, FRE: Flesch Reading Ease

Discussion and Conclusion

This study compared the performance of GPT-4.o and Gemini-1.5-Flash in answering cardiology-related multiple-choice questions. GPT-4.o demonstrated a superior performance, particularly with a correct response rate exceeding the 60% threshold that is typically considered a passing grade for many exams. In contrast, Gemini-1.5-Flash’s correct response rate of 58.6% indicated a failure to meet this standard.

In terms of human performance, it is well-documented that CS-MCQs are designed to assess higher-order thinking skills such as analysis, problem-solving, and knowledge integration. These questions challenge students to think beyond isolated medical facts, instead encouraging a holistic view of the patient[23,24]. Case and Swanson (2002) have noted the particular importance of case-based clusters in PBL settings, as they test the practical application of knowledge[25].

In our study, both CS-MCQs and stand-alone questions (non-case questions) were used to assess the models’ current level of learning. Stand-alone questions, which do not involve case presentations, mainly test basic recall of medical knowledge. Bhayana et al. found that LLMs performed well on questions requiring lower levels of cognitive processing, such as basic knowledge recall[26]. In our findings, Gemini significantly outperformed GPT-4.o in non-case questions, which could suggest that Gemini's training may have been based on a similar dataset and also that Gemini may not have been able to effectively apply the information it learned to more complex formats such as case-based scenarios.

However, GPT-4.o excelled in case-based questions, indicating its superior ability to apply knowledge beyond mere recall. When analyzing question subgroups, Gemini-1.5-Flash performed consistently across all categories, while GPT-4.o struggled with questions that included images but lacked case details. This suggests that GPT-4.o's training may have emphasized textual over visual elements, with additional textual details facilitating better performance. In contrast, previous studies have shown ChatGPT-4’s effectiveness in responding to clinical image-based questions. ChatGPT has been particularly useful in diagnostic decision-making within radiology[27,28], although in our study only ECG images were used, rather than a broader range of radiological images. Group 2 (with an image, without a case presentation) showed the lowest performance from GPT-4.o, indicating that the presence of an image without a case presentation hindered correct responses.

The FKGL and the FRE scores consider the number of sentences and words to determine a text’s reading level. Our study results revealed that the responses of both GPT-4.o and Gemini-1.5-Flash were at the level of college graduate. Consistent with our results, in a previous study, ChatGPT’s FKGL and FRE scores indicated a hard reading level appropriate for only 33% of adults and those with a college education[29]. Furthermore, the texts produced by ChatGPT were harder than those from Bard, Gemini's predecessor[30]. Our results show similar characteristics, as Gemini-1.5-Flash's FKGL was significantly lower than GPT-4.o's, and Gemini-1.5-Flash's score was higher than GPT-4.o. Conversely, Atkinson et al. identified that although ChatGPT's responses were consistently accurate, they were somewhat superficial and corresponded to the knowledge level of a trainee[31]. Rizwan et al. reported that if healthcare information was not sufficient, reaching into consistent conclusion based on ChatGPT has proved to be an efficient and effective tool both academically and in clinical setups[32].

AI models like GPT-4.o and Gemini-1.5-Flash generate responses based on patterns in their training data, which means that their answers are probabilistic rather than absolute. While their output can seem authoritative, AI models can produce misinformation or misunderstanding due to their lack of deep, principle-based medical reasoning.

In the field of cardiology, AI has made strides, including more accurate predictions of myocardial infarction (MI) risk than traditional methods and successfully passing the European Exam in Core Cardiology (EECC)[33,34]. However, it remains unclear to what extent these models base on specific medical guidelines such as American College of Cardiology (ACC), American Heart Association (AHA) and European Society of Cardiology (ESC) standards. Also, studies about ECG interpretation performance of AI are present[35-37]. In a previous study it was found that the limited accuracy and consistency of GPT-4 and Gemini suggest that their current use in clinical ECG interpretation is risky[36]. Further research is needed to quantitatively assess AI variability in clinical settings to better understand its reliability.

The ethical implications of LLMs in medicine are significant. In cases where AI provides erroneous advice leading to negative outcomes, questions of liability arise. Notably, Gemini-1.5-Flash warns users that it cannot offer medical advice, while GPT-4.o provides direct responses. Additionally, concerns over data privacy and security are critical, as current LLMs store information on their servers. Therefore, AI applications must warn their users that personal information may be uploaded anonymously. For AI to be fully integrated into clinical practice, it must be able to handle personal patient data securely[38]. AI systems are also subject to biases from their training data, potentially leading to outdated information, unequal care, or even discrimination[39,40].

On the other hand, the integration of artificial intelligence into emergency room practice is of critical importance in many countries due to reasons such as the long waiting times of patients in emergency rooms, the lack of doctors from all departments in emergency rooms 24 hours a day, physician shortage and extraordinary conditions such as the COVID-19 pandemic or earthquakes, where emergency rooms will be visited far beyond their capacity. In the future, as artificial intelligence technology develops, the initial evaluation can be made more comprehensive by using artificial intelligence applications, at least while waiting to reach the relevant branch physician in emergency room consultations.

Study Limitations

This study evaluated GPT-4.o and Gemini-1.5-Flash solely in the English language. While both models are capable of communicating in multiple languages, their performance depends on the quality and amount of training data available in a specific language. Since English is the most common language in AI training, performance in other languages may be lower. Further research is required to evaluate LLM performance across different languages. Additionally, the number of questions in this study was limited. Future studies should include much more questions with diverse question types. Also, in our study, questions were posed to GPT-4o and Gemini-1.5-Flash only once. Therefore, the study does not reflect how performance might vary in repeated questioning. It also does not allow for evaluation of the models' consistency in their responses.

Conclusion:

Both models demonstrated an ability to comprehend the scenarios presented and provided reasonable answers. Despite the limitations and ethical concerns surrounding the use of AI in medicine, it is essential for physicians to remain engaged with ongoing AI research and support its responsible development. The integration of AI could potentially elevate the standards of medical care, education, research, and clinical decision-making in the future. However, AI should not be seen as a replacement for critical thinking, creativity, and innovation—skills that remain uniquely human and are crucial in the medical field.

Acknowledgements: I would like to thank Medscape executives and the editorial coordinator Kennedy Maklos (Schaff), MHCDS for allowing us to use case presentations and ECG questions that are available to healthcare professionals on Medscape's online platform. The manuscript was prepared entirely with no direct assistance from generative artificial intelligence or similar technologies.

Researcher Contribution Statement:
Idea and design: S.G.P., D.S..; Data collection and processing: S.G.P, V.A.D., C.A., I.I.K. Analysis and interpretation of data: D.S., S.G.P.; Writing of significant parts of the article: S.G.P., D.S, V.A.D.

Support and Acknowledgement Statement:
No financial support

Conflict of Interest Statement:
The authors of the article have no conflict of interest declarations.
Ethics Committee Approval Information:
This study does not include any human or animal participants. Ethics committee approval is not applicable like the similar studies in literature.

References

1. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ. 2023;20(56). doi:10.1186/s41239-023-00416-7
2. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inform Manag. 2023;71:102642. doi:10.1016/j.ijinfomgt.2023.102642
3. Yenduri G. GPT (Generative Pre-Trained Transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access. 2024;12:1-36. doi:10.1109/ACCESS.2024.3389497
4. Hadi MU, Al-Tashi Q, Qureshi R, et al. Large language models: a comprehensive survey of applications, challenges, limitations, and future prospects. Authorea. Preprint. 2023.
5. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. 2023. doi:10.21203/rs.3.rs-2924050/v1
6. Saka A, Taiwo R, Saka N, et al. GPT models in construction industry: opportunities, limitations, and a use case validation. Dev Built Environ. 2024;17:100300. doi:10.1016/j.dibe.2023.100300
7. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial intelligence-powered drug discovery. Nat Mach Intell. 2022;4(3):189-191. doi:10.1038/s42256-022-00480-0
8. OpenAI. GPT-4 Technical Report. 2023. Available at: https://cdn.openai.com/papers/gpt-4.pdf. Accessed June 16, 2025.
9. OpenAI. Introducing GPT-4o and more tools to ChatGPT free users. 2024. Available at: https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/. Accessed June 16, 2025.
10. Chen CH, Hsieh KY, Huang KE, Lai HY. Comparing vision-capable models, GPT-4 and Gemini, with GPT-3.5 on Taiwan's pulmonologist exam. Cureus. 2024;16(8):e67641. doi:10.7759/cureus.67641
11. Masanneck L, Schmidt L, Seifert A, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. J Med Internet Res. 2024;26:e53297. doi:10.2196/53297
12. Builoff V, Shanbhag A, Miller RJ, et al. Evaluating AI proficiency in nuclear cardiology: large language models take on the board preparation exam. medRxiv. Preprint. 2024. doi:10.1101/2024.07.16.24310297
13. Botross M, Mohammadi SO, Montgomery K, Crawford C. Performance of Google's artificial intelligence chatbot "Bard" (now "Gemini") on ophthalmology board exam practice questions. Cureus. 2024;16(3):e57348. doi:10.7759/cureus.57348
14. Khan MP, O'Sullivan ED. A comparison of the diagnostic ability of large language models in challenging clinical cases. Front Artif Intell. 2024;7:1379297. doi:10.3389/frai.2024.1379297
15. Hirosawa T, Harada Y, Mizuta K, et al. Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians: experimental study for diagnostic cases. JMIR Form Res. 2024;8:e59267. doi:10.2196/59267
16. Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large language models for intraoperative decision support in plastic surgery: a comparison between ChatGPT-4 and Gemini. Medicina (Kaunas). 2024;60(6):957. doi:10.3390/medicina60060957
17. Rush R. Assessing readability: formulas and alternatives. Read Teach. 1984;39(3):274-283.
18. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233. doi:10.1037/h0057532
19. Medscape. Case Challenges. Available at: https://reference.medscape.com/features/casechallenges?icd=login_success_email_match_norm. Accessed September 6, 2024.
20. Medscape. Home Page. Available at: https://www.medscape.com/index/section_60_0. Accessed September 6, 2024.
21. Readable. Flesch Reading Ease and the Flesch Kincaid Grade Level. Available at: https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/. Accessed June 16, 2025.
22. Klare GR. The measurement of readability: useful information for communicators. ACM J Comput Doc. 2000;24:107-121.
23. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? BMC Med Educ. 2007;7:49. doi:10.1186/1472-6920-7-49
24. Hays RB, Coventry P, Wilcock D, Hartley K. Short and long multiple-choice question stems in a primary care oriented undergraduate medical curriculum. Educ Prim Care. 2009;20(3):173-177.
25. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd ed. Philadelphia: National Board of Medical Examiners; 2002.
26. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023;307(5):e230582. doi:10.1148/radiol.230582
27. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. 2023. doi:10.1101/2023.02.02.23285399
28. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. Radiology. 2023;307:e230171.
29. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina. 2023;7:862-868.
30. Al-Sharif EM, Penteado RC, Dib El Jalbout N, et al. Evaluating the accuracy of ChatGPT and Google Bard in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. Ophthalmic Plast Reconstr Surg. 2024;40:303-311.
31. Atkinson CJ, Seth I, Xie Y, et al. Artificial intelligence language model performance for rapid intraoperative queries in plastic surgery: ChatGPT and the deep inferior epigastric perforator flap. J Clin Med. 2024;13:900.
32. Rizwan A, Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. Cureus. 2023;15(8):e43106. doi:10.7759/cureus.43106
33. Mahendiran T, Thanou D, Senouf O, et al. Deep learning-based prediction of future myocardial infarction using invasive coronary angiography: a feasibility study. Open Heart. 2023;10:e002237.
34. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial

intelligence success story? Eur Heart J Digit Health. 2023;4:279-281.

35. Herman R, Kisova T, Belmonte M, et al. Artificial intelligence-powered electrocardiogram detecting culprit vessel blood flow abnormality: AI-ECG TIMI study design and rationale. J Soc Cardiovasc Angiogr Interv. 2025;4(3Part B):102494. doi:10.1016/j.jscai.2024.102494
36. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. Am J Emerg Med. 2024;84:68-73. doi:10.1016/j.ajem.2024.07.043
37. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. J Cardiovasc Dev Dis. 2023;10(4):175. doi:10.3390/jcdd10040175
38. Yuan J, Tang R, Jiang X, Hu H. Large language models for healthcare data augmentation: an example on patient-trial matching. AMIA Annu Symp Proc. 2023;2023:1324-1333.
39. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of COVID-19 healthcare? BMJ. 2021;372:n304.
40. Zaidi D, Miller T. Implicit bias and machine learning in health care. South Med J. 2023;116:62-64.