Thorac Res Pract. **DOI:** 10.4274/ThoracResPract.2026.2026-1-2

Original Article

Comparison of AI-based Chatbot Performance in Analyzing Clinical Scenarios versus Medical Residents: A Novel Approach in Chest Diseases Education

Mehmet Hakan Bilgin[1], Hamit Hakan Alp[2]

[1]Department of Chest Diseases, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye
[2]Department of Biochemistry, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye

Cite this article as: Bilgin MH, Alp HH. Comparison of AI-based chatbot performance in analyzing clinical scenarios versus medical residents: a novel approach in chest diseases education. *Thorac Res Pract.* [Epub Ahead of Print]

ABSTRACT

OBJECTIVE: Rapid advancements in artificial intelligence (AI) technologies offer new opportunities in medical education. The aim of this study is to compare the performance of large language models, specifically ChatGPT-4 and Gemini, in analyzing clinical scenarios with that of chest diseases research assistants (residents), and to evaluate their potential roles in medical education.

MATERIAL AND METHODS: This cross-sectional, comparative study included 28 resident physicians working in the department of chest diseases at a tertiary-care university hospital. Four clinical scenarios involving diagnoses of massive pulmonary embolism, chronic obstructive pulmonary disease, asthma, and severe pneumonia/sepsis were presented to both participants and AI models (ChatGPT-4 and Gemini). Responses were scored by blinded experts based on current guidelines (Global Initiative for Chronic Obstructive Lung Disease, Global Initiative for Asthma, American Thoracic Society).

RESULTS: AI models achieved significantly higher scores than residents, particularly on structured questions requiring theoretical knowledge, classification skills, and the listing of contraindications ($P < 0.05$). However, it was observed that residents achieved success levels similar to those of AI models in situations requiring emergency intervention (e.g., shock management) through practical, results-oriented approaches. While AI models provided a broader spectrum in differential diagnosis, residents preferred "telegraphic" and practice-oriented responses.

CONCLUSION: ChatGPT and Gemini have significant potential as clinical decision-support systems and educational assistants. However, rather than replacing human factors in clinical reasoning and emergency management, they should be positioned as complementary tools that accelerate physicians' access to theoretical knowledge.

KEYWORDS: Education, medical, artificial intelligence, diagnosis, differential, clinical reasoning, decision support systems, clinical, humans

Received: 05.01.2026 ***Revision Requested:*** 09.02.2026 ***Last Revision Received:*** 11.02.2026

Accepted: 23.02.2026 ***Epub:*** 14.04.2026

INTRODUCTION

Artificial intelligence (AI), particularly chatbots with natural language processing capabilities, promises transformative changes in healthcare services and medical education. As stated by Davenport and Kalakota[1], the potential of AI to provide data analysis and diagnostic support in healthcare spans a wide range, from reducing administrative workloads to complex clinical decisions. With the introduction of Generative Pre-trained Transformer models like ChatGPT, a paradigm shift is occurring in how patients access information and in physicians' decision-making processes.[2]

Medical education is evolving from the traditional master–apprentice relationship to digitally supported learning models. In a widely cited study by Kung et al.[3], it was demonstrated that ChatGPT could pass the United States Medical Licensing Examination (USMLE) without any specific training. This suggests that AI is capable not only of storing information but also of demonstrating clinical reasoning abilities. Similarly, Singhal et al.[4] reported that Med-PaLM and Gemini models

Corresponding author: Mehmet Hakan Bilgin, MD, e-mail: medicaldrhakan@gmail.com, mehmethakanbilgin@yyu.edu.tr

developed by Google performed at near-expert levels in medical questions.

However, the reliability of these models in clinical practice remains a subject of debate. Huang et al.[5], in their study comparing family medicine residents with AI models, noted that while AI showed high success in structured examinations, human reasoning remains critical in complex cases involving "diagnostic uncertainty." Studies in specific branches such as ophthalmology and cardiology also indicate that while AI's capacity to provide information is promising, the risk of "hallucination" (generating false information) must not be ignored.[6,7]

The practice of pulmonology (chest diseases) is a complex field in which emergency decision-making (e.g., pulmonary embolism, sepsis) and chronic disease management [chronic obstructive pulmonary disease (COPD), asthma] are intertwined. This study aims to reveal the place of these tools in residency training by using concrete data to compare the performance of AI models (ChatGPT and Gemini) in clinical scenarios with that of residents actively working in the field.

MATERIAL AND METHODS

Study Design and Ethical Approval

This study is a cross-sectional simulation conducted at the department of chest diseases of a tertiary care university hospital. Approval for the study was obtained from the Van Yüzüncü Yıl University Ethics Committee (decision dated: 31.10.2025 and numbered: 34319). The study was conducted in accordance with the principles of the Declaration of Helsinki.

Participants

The study sample consisted of 28 resident physicians who were receiving specialization training in the relevant clinic and who volunteered to participate. ChatGPT-4 developed by OpenAI and Gemini Advanced (utilizing the Ultra 1.0 model; Google, Mountain View, CA).

Main Points

- Artificial intelligence (AI) models (ChatGPT-4 and Gemini) demonstrated superior performance compared with residents in assessments of theoretical knowledge and guideline-based classification tasks.
- Residents achieved success rates similar to those of AI models in emergency scenarios and exhibited more practical, action-oriented approaches.
- AI models show high potential as educational support tools in residency training, but they require human oversight for clinical reasoning and ethical decision-making.

Data Collection Tools (Scenarios)

Four standard clinical scenarios, prepared by expert faculty members, were used to assess participants' knowledge and clinical approach.

- Scenario 1 (Vascular Emergency): A case of massive pulmonary embolism (MPE) in the post-op period presenting with shock.
- Scenario 2 (Obstructive Lung Disease): A case of COPD presenting with exacerbation.
- Scenario 3 (Chronic Airway): An asthma case presenting in an outpatient setting.
- Scenario 4 (Infection): A young patient with severe pneumonia and sepsis.

For each scenario, 5–6 open-ended questions were asked regarding diagnosis, differential diagnosis, test requests, treatment plans, and discharge criteria. To ensure standardization, a specific prompt was used to query the AI models. The prompt was structured as follows: "Act as a senior chest diseases specialist. Evaluate the following clinical case according to current guidelines [Global Initiative for Chronic Obstructive Lung Disease (GOLD), Global Initiative for Asthma (GINA), American Thoracic Society] and answer the questions step by step." The clinical scenarios were presented to both AI models and residents in the same order. To prevent bias during the evaluation, the outputs from the AI models and the handwritten responses from the residents were transcribed into a uniform digital format (Times New Roman, 12 pt) to ensure effective blinding. The evaluators were unaware of whether a response belonged to an AI or a resident.

The residents completed the evaluation in a supervised classroom environment as a "closed-book" exam, without access to external resources, internet, or mobile devices. In contrast, AI models utilized their internal databases. This difference in access to information is acknowledged as a limitation of the study.

Implementation and Evaluation

Residents answered the questions under supervision without using external resources. The same questions were presented to the AI models without employing prompt-engineering techniques. Instead, a single neutral instruction was used to initiate the session: "Act as an expert chest diseases specialist and answer the following clinical scenario questions based on current medical guidelines." Subsequently, the scenario texts were entered directly. The responses were evaluated by two independent blinded experts using a standardized scoring system based on current guidelines. Each scenario was evaluated based on five key components: (1) Correct Diagnosis, (2) Appropriate Diagnostic Tests, (3) Treatment Strategy, (4) Dosage and Duration Accuracy, and (5) Follow-up Plan. Each component was scored on a scale of 0 to 2 (0: Incorrect/Missing, 1: Partially Correct, 2: Fully Correct/Guideline-Compliant), resulting in a maximum total score of 10 points per scenario. The average score given by the two experts was recorded as the final score.

Statistical Analysis

Categorical data were summarized as numbers and percentages, and continuous variables that were not normally distributed were presented as medians and interquartile ranges. Categorical data were compared between groups using Pearson's χ^2 test and Fisher's exact test. Because the numerical variables were not normally distributed, the Mann-Whitney U test, a non-parametric test, was used to compare continuous variables.

RESULTS

The responses of 28 resident physicians were compared with those of the ChatGPT-4 and Gemini models.

Score Comparison

Overall, AI models achieved significantly higher scores than the resident group ($P < 0.05$). However, when analyzed by scenario, the gap between resident success scores and AI scores narrowed in scenarios requiring emergency management, such as MPE. The detailed score distribution is presented in Table 1.

Diagnostic Accuracy and Theoretical Knowledge

In all scenarios, both the resident group and the AI models demonstrated nearly 100% success in making the "correct diagnosis." However, significant differences were detected in the level of detail of treatment protocols and in the classification questions (Table 2).

• COPD Scenario: While 75% of the residents performed GOLD staging (groups A, B, E) correctly, AI models responded with 100% accuracy and provided references to current guidelines.

• MPE: In the section questioning thrombolytic treatment contraindications, residents were able to list an average of 1.8±0.6 items (usually "active bleeding" and "surgery"), whereas AI models listed an average of 8 items (all absolute and relative contraindications).

Clinical Approach Differences (Qualitative Analysis)

It was observed that resident responses were shorter, more concise, and "action-oriented." For example, in the pneumonia scenario, while residents directly wrote the antibiotic from the hospital protocol (e.g., "Start Pip-Tazo"), AI models listed all alternatives in the guidelines (Quinolones, Macrolide combinations, etc.). Residents demonstrated high proficiency in the practical application of the qSOFA and CURB-65 scoring systems. The qualitative differences between human and AI responses are summarized in Table 3.

DISCUSSION

This study found that large language models (LLMs) such as ChatGPT-4 and Gemini performed better than residents receiving specialization training in solving clinical scenarios in the field of chest diseases, particularly in terms of theoretical knowledge and mastery of structured guidelines. Our findings reveal that LLMs can be powerful assistants in medical education and decision-support processes, but they cannot yet fully replace the "human factor" in practical decision-making.

Regarding theoretical knowledge and guideline mastery, AI models provided more comprehensive responses than residents, especially in questions requiring "recall," such as listing COPD staging and thrombolytic treatment contraindications. This finding supports Kung et al.'s[3] work on ChatGPT's success in the USMLE exam and Gilson et al.'s[8] studies on knowledge assessment in medical education. AI models drew an "ideal resident" profile by presenting current classifications in GOLD and GINA guidelines without error. Nori et al.[9], in their study on the medical proficiency of GPT-4, suggested that the model does not merely memorize but can also synthesize complex

Table 1. Comparison of scenario-based success scores between resident physicians and AI models

Scenario	Resident physicians (n = 28) (mean score ± SD)	ChatGPT-4 (score)	Gemini (score)	*P* value*
Scenario 1: massive PE	82.5±12.4	95.0	98.0	**0.042**
Scenario 2: COPD exacerbation	78.0±14.1	92.0	94.0	**0.038**
Scenario 3: asthma	88.0±9.5	96.0	96.0	0.112
Scenario 4: severe pneumonia/sepsis	85.5±11.2	95.0	97.0	0.085
General average	83.5±11.8	94.5	96.25	**<0.05**

*Mann-Whitney U test (significance of difference between resident mean and AI models)
AI: artificial intelligence, PE: pulmonary embolism, COPD: chronic obstructive pulmonary disease, SD: standard deviation

Table 2. Correct response rates according to evaluation criteria

Evaluation criteria	Resident physicians (correct response %)	AI models (correct response %)	Statistical significance *(P)*
Diagnostic accuracy	96.4%	100%	>0.05
Scope of differential diagnosis	65.0%	100%	**<0.001**
Compliance with guidelines (GOLD/GINA)	71.4%	100%	**<0.05**
Listing treatment contraindications	60.7%	100%	**<0.001**
Emergency management/practical approach	92.8%	90.0%	>0.05

GOLD: Global Initiative for Chronic Obstructive Lung Disease, GINA: Global Initiative for Asthma

clinical data; the high success rate in asthma/COPD differential diagnosis in our study confirms this synthesis ability.

However, a distinct difference emerged between clinical reasoning and the "human touch". When the responses of our residents were examined, they exhibited an "action-oriented" approach rather than focusing on theoretical details. For example, in emergency scenarios such as MPE, residents focused on stabilizing vital functions, whereas AI provided a more didactic breakdown. This fundamental difference in approach — human pragmatism versus AI exhaustiveness — is visually summarized in Figure 1, which contrasts the action-oriented nature of residents with the information-centric nature of AI.

As stated by Sabaner et al.[6], although medical chatbots are successful in providing information, they remain insufficient in areas requiring empathy, intuition, and ethical responsibility, which are the cornerstones of patient management. As emphasized by Lee et al.[10] in the New England Journal of Medicine, while GPT-4 shows remarkable capabilities, it requires human oversight to mitigate risks such as hallucinations; therefore, AI can currently serve only as a "co-pilot" rather than a replacement for physicians. Furthermore, as noted by Rao et al.[11], since AI models cannot know local hospital conditions and resource constraints, residents' treatment arrangements based on available facilities (e.g., antibiotic stock on hand) represent a more realistic approach.

Another critical aspect to consider is the risk of "hallucination" and the issue of reliability. One of the biggest handicaps of AI models is the condition defined in the literature as hallucination.[12] No significant hallucinations were detected in the AI models used in our study, because the scenarios were based on standard medical guidelines. This suggests, as indicated by Johnson et al.[13], that AI reliability is high in well-defined diseases. However, considering the warning by Shaw et al.[14], it is critically important for students not to develop "automation bias" (over-reliance) when using these tools in education.

These findings have significant educational implications. Our findings indicate that AI can serve as an "information provider" in chest disease education. Just as Jang and Kim[15] emphasized the success of video-based learning models, LLM-based interactive scenarios can reinforce learning in residency

Table 3. Qualitative comparison of human (resident) and AI responses

Feature	Resident physicians (human)	AI models (ChatGPT/Gemini)
Response style	Result-oriented, concise, "telegraphic"	Structured, detailed, didactic
Information source	Clinical experience + theoretical knowledge	Extensive database + guidelines
Differential diagnosis	Focuses on the most likely 2–3 diagnoses	Lists all possible diagnoses (including rare ones)
Treatment approach	Based on hospital facilities/protocols	Presents all options in academic guidelines
Most common error	Omission in listing questions (forgetting items)	Rarely presenting contextually detached suggestions

AI: artificial intelligence

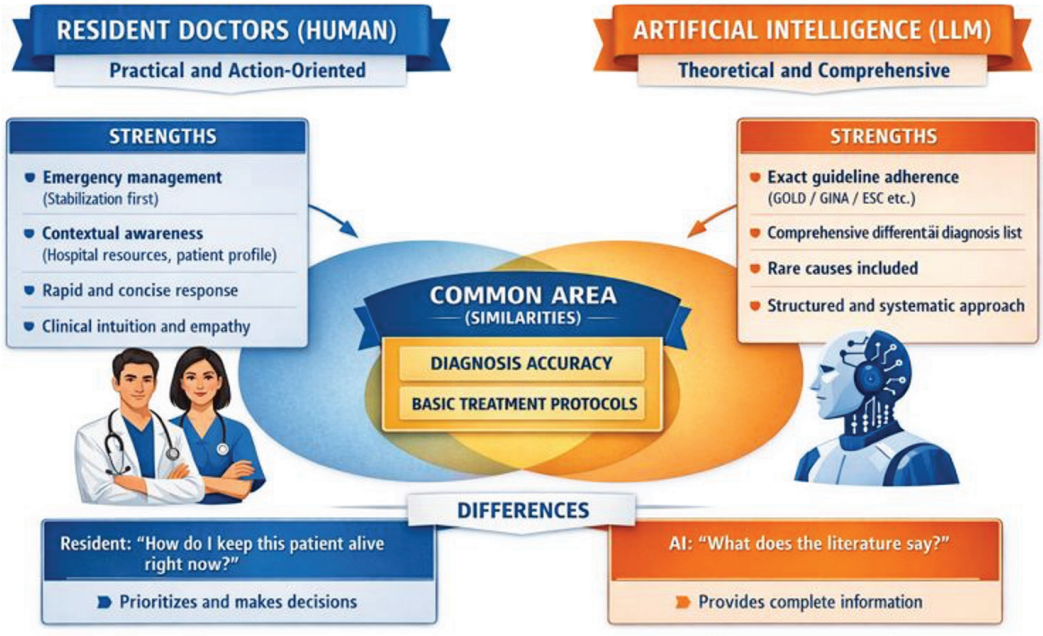

Figure 1. Conceptual comparison of clinical approaches: while resident physicians focus on "practical action" and "contextual awareness," AI models excel in "theoretical knowledge" and "structured listing." Both groups intersect at "diagnostic accuracy"

AI: artificial intelligence, LLM: large language model, GOLD: Global Initiative for Chronic Obstructive Lung Disease, GINA: Global Initiative for Asthma, ATS: American Thoracic Society, ESC: European Society of Cardiology

training. As highlighted in the systematic review by Tozsin et al.[16], AI applications have the potential to significantly contribute to medical education by providing personalized feedback and increasing student engagement. Therefore, integrating AI tools into the curriculum as a supplementary resource, rather than a replacement, appears to be the most beneficial approach. These results suggest that residency training should evolve to focus less on rote memorization and more on practical bedside skills and crisis management, because AI can effectively support theoretical queries.

Study Limitations

This study has several limitations. First, the study was conducted at a single center and included a relatively small sample of resident physicians (n = 28). This constraint may restrict the generalizability of our findings to other institutions or healthcare systems with different residency training curricula. Second, the scenarios were text-based, and visual data (such as chest X-rays or computed tomography scans) were not directly analyzed by the AI models; direct analysis of such visual data is an integral part of daily pulmonology practice. Finally, resident physicians responded under potential examination-related stress and time constraints, whereas AI models were exempt from such psychological pressures and fatigue, a factor that might have influenced the performance gap. Another limitation is the inherent disparity in testing conditions: residents answered without access to external resources, whereas AI models had instant access to vast datasets. This "unfair comparison" highlights the difference between human cognitive recall and AI data retrieval.

CONCLUSION

ChatGPT and Gemini demonstrate high accuracy in diagnosis and treatment planning for clinical scenarios involving chest diseases. AI is more effective than human memory at recalling theoretical knowledge and guideline details. However, resident physicians demonstrate their clinical competence through practical and goal-oriented approaches in emergency situations. AI models should be used as powerful educational support tools capable of rapidly addressing residents' knowledge gaps, but decision-making must always remain under the control of the physician.

Ethics

Ethics Committee Approval: Approval for the study was obtained from the Van Yüzüncü Yıl University Ethics Committee (decision dated: 31.10.2025 and numbered: 34319). The study was conducted in accordance with the principles of the Declaration of Helsinki.

Informed Consent: Written informed consent was obtained from all participants.

Footnotes

Authorship Contributions

Surgical and Medical Practices: M.H.B., Concept: M.H.B., H.H.A., Design: M.H.B., H.H.A., Data Collection or Processing: M.H.B., H.H.A., Analysis or Interpretation: M.H.B., H.H.A., Literature Search: M.H.B., H.H.A., Writing: M.H.B., H.H.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* 2019;6(2):94-98. **[Crossref]**
2. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *NCI Cancer Spectr.* 2023;7(2):pkad010. **[Crossref]**
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* 2023;2(2):e0000198. **[Crossref]**
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172-180. **[Crossref]**
5. Huang RS, Benour A, Kemppainen J, Leung FH. The future of AI clinicians: assessing the modern standard of chatbots and their approach to diagnostic uncertainty. *BMC Medical Educ.* 2024;24(1):1133. **[Crossref]**
6. Sabaner MC, Anguita R, Antaki F, et al. Opportunities and challenges of chatbots in ophthalmology: a narrative review. *J Pers Med.* 2024;14(12):1165. **[Crossref]**
7. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA.* 2023;329(10):842-844. **[Crossref]**
8. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312. **[Crossref]**
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv.* 2023. **[Crossref]**
10. 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):1233-1239. **[Crossref]**
11. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *MedRxiv.* 2023:2023.02. **[Crossref]**
12. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys.* 2023;55(12):248. **[Crossref]**
13. 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:rs.3.rs-2566942. **[Crossref]**
14. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. *Med Sci Educ.* 2025:35(3):1803-1816. **[Crossref]**
15. Jang HW, Kim KJ. Use of online clinical videos for clinical skills training for medical students: benefits and challenges. *BMC Med Educ.* 2014;14:56. **[Crossref]**
16. Tozsin A, Ucmak H, Soyturk S, et al. The role of artificial intelligence in medical education: a systematic review. *Surg Innov.* 2024;31(4):415-423. **[Crossref]**