Official Journal of Erciyes University Faculty of Medicine

DOI: 10.14744/cpr.2026.71836
J Clin Pract Res 2026;48(2):145–152

Comparison of Intern Doctors and ChatGPT in Emergency Cases Assessment

Yusuf Kantar,[1] Melih İmamoğlu,[2] Emre Bülbül,[3] Serhat Hayme,[4] Arif Onur Eden,[1] Yasin Bilgin,[1] Fatih Mehmet Sari[1]

[1]Department of Emergency Medicine, Erzincan Binali Yıldırım University, Faculty of Medicine, Erzincan, Türkiye
[2]Department of Emergency Medicine, Karadeniz Technical University, Faculty of Medicine, Trabzon, Türkiye
[3]Department of Emergency Medicine, Erciyes University, Faculty of Medicine, Kayseri, Türkiye
[4]Department of Biostatistics and Medical Informatics Erzincan Binali Yıldırım University, Faculty of Medicine, Erzincan, Türkiye

Cite this article as:
Kantar Y, İmamoğlu M, Bülbül E, Hayme S, Eden AO, Bilgin Y, Sari FM. Comparison of Intern Doctors and ChatGPT in Emergency Cases Assessment. J Clin Pract Res 2026;48(2):145–152.

Address for correspondence:
Yusuf Kantar.
Department of Emergency Medicine, Erzincan Binali Yıldırım University, Faculty of Medicine, Erzincan, Türkiye
Phone: +90 541 927 58 17
E-mail:
dryusufkantar@gmail.com

Submitted: 17.12.2025
Revised: 05.03.2026
Accepted: 26.03.2026
Available Online: 20.04.2026

Erciyes University Faculty of Medicine Publications - Available online at www.jcpres.com

ABSTRACT

Objective: Accurate and timely diagnosis in emergency departments is crucial due to the high patient volume and time-sensitive nature of care. Intern doctors, who are nearing the completion of medical school, frequently work in emergency departments in many countries. However, after graduation, physicians are often expected to assume critical patient care responsibilities despite limited experience. Artificial intelligence models can quickly analyze patient data and generate diagnoses, thus assisting inexperienced physicians in enhancing diagnostic accuracy. This study aims to evaluate the diagnostic performance of ChatGPT-4 in emergency department case scenarios and compare its accuracy with that of intern doctors.

Materials and Methods: This study involved intern doctors participating in the internship program during the 2024–2025 academic year. A total of 36 case-based questions, categorized by difficulty level, were administered to 155 interns and subsequently presented to artificial intelligence. Descriptive statistics were used to summarize the data, and a one-sample t-test was conducted to compare the diagnostic accuracy between intern doctors and ChatGPT. Statistical significance was set at $p<0.05$.

Results: Intern doctors achieved an overall correct response rate of 58.3%, while ChatGPT achieved a rate of 97.2%. A statistically significant, moderate negative correlation was found between question difficulty and interns' performance ($r=-0.684$; $p<0.001$), indicating decreased accuracy as question difficulty increased. ChatGPT consistently demonstrated significantly higher performance across all difficulty levels.

Conclusion: ChatGPT-4 may serve as a valuable diagnostic support tool in emergency departments, particularly for newly graduated physicians with limited clinical experience.

Keywords: Artificial intelligence, ChatGPT, emergency department, intern doctors, medical education.

INTRODUCTION

Medical education is a long and demanding process that requires sustained effort and perseverance. Its final phase is the internship period, during which medical students serve as intern doctors (IDs). In contrast to the predominantly theoretical training received throughout their academic years, the internship provides a structured opportunity for experiential learning, allowing students to apply their knowledge in real clinical settings.[1] Throughout the year, IDs rotate through multiple clinical departments, including emergency medicine. The emergency medicine rotation, however, has several distinguishing features compared with other specialties. Patients typically present to the emergency department (ED) with acute and potentially life-threatening conditions, requiring clinicians to rapidly generate differential diagnoses and initiate timely management. This environment necessitates the development of strong clinical decision-making skills under pressure.[2]

Secondly, and perhaps more importantly, most newly graduated physicians are required to work in EDs for a designated period before selecting a specialty. This requirement stems from the mandatory service obligation implemented by the Ministry of Health in our country. The primary assignments for this compulsory service are EDs and prehospital ambulance services. As a result, newly graduated doctors—often with limited clinical experience—are expected to provide critical patient care immediately after completing medical school. Therefore, it is essential that intern doctors receive comprehensive, high-quality training to adequately prepare them for these responsibilities.[3]

The artificial intelligence (AI) program utilized in our study is ChatGPT, a GPT-based AI model developed by OpenAI, which has made substantial progress in natural language processing.[4] This model is distinguished by its capacity to deliver a human-like experience in text-based communication. Although AI systems have not yet been integrated into routine clinical practice, they have been the focus of experimental research for several years. Their ability to accurately interpret user inputs and generate rapid, coherent outputs indicates strong potential for future applications in the healthcare domain.[5]

In the context of the ED, despite the rising volume of patient visits, several studies have shown that AI can be implemented as early as the triage stage to facilitate the efficient progression of diagnostic and therapeutic processes.[6] Furthermore, when data from patients diagnosed and admitted by ED residents were input into AI programs, the AI generated diagnostic suggestions consistent with those of the physicians.[7] Studies have also indicated that AI can support the diagnostic process, particularly in the interpretation of medical tests. It has shown high efficacy in analyzing radiological images and electrocardiography (ECG) results.[8,9]

KEY MESSAGES

- ChatGPT can serve as a valuable guide for newly graduated doctors who are theoretically well-equipped, helping them apply their knowledge until they gain sufficient practical experience.
- Although artificial intelligence appears successful in case scenarios, presenting these cases to it requires a certain level of medical expertise and knowledge.
- In critical situations, such as patient care, artificial intelligence cannot be given full responsibility. It can only serve as a guide for healthcare professionals.

In this context, it can be confidently asserted that integrating AI-driven technologies into time-sensitive clinical environments such as EDs will enhance patient care and support inexperienced physicians in the near future. By facilitating more accurate and timely diagnoses in critical and complex cases, AI not only promotes public health and the efficient use of healthcare resources but also helps safeguard physicians against malpractice claims.

This study aims to assess ChatGPT's ability to accurately predict diagnoses in ED cases based on provided clinical findings, in comparison with IDs.

MATERIALS AND METHODS

This study is a single-center, cross-sectional observational study. The study population included intern doctors (IDs) who were enrolled in the emergency department (ED) rotation during the 2024–2025 academic year and consented to participate. Ethical approval was obtained from the Erzincan Binali Yıldırım University Non-Interventional Clinical Research Ethics Committee (Approval Number: 2024-13/05, Date: 03.10.2024). This study was conducted in accordance with the principles of the Declaration of Helsinki.

The hospital where the study was conducted is a university hospital, serving as the only public hospital in the city and providing care to a population of approximately 500,000, including patients from surrounding provinces and districts. The ED receives between 1,500 and 2,000 patient visits daily.

The questions were developed by a panel of five faculty members from different medical schools, each with at least five years of professional experience in emergency medicine. This expert panel (EP) designed 36 multiple-choice questions based on scenarios relevant to ED practice. The development

process—including question formulation, arrangement of answer options, and determination of the correct answer—was guided by current literature, clinical guidelines, and the panelists' professional experience. For each question, a single correct answer was established as the gold standard (GS). Participants were subsequently asked to select the "most likely diagnosis" for each case. Their responses were then compared with the GS answers.

Twelve questions were categorized as easy, 12 as moderate, and 12 as difficult. The difficulty level of each scenario was determined by the EP based on consensus, taking into account factors such as presenting complaints, physical examination findings, diagnostic methods used, and the interpretability of test results in relation to the clinical presentation. Participants answered the questions without knowledge of their assigned difficulty levels, ensuring blinding with respect to question difficulty.

To ensure the validity of the difficulty classification, inter-rater reliability among the EPs who categorized the questions was assessed using Fleiss' kappa, an appropriate method for evaluating agreement among multiple raters. Each specialist independently assigned all 36 questions to the easy, moderate, or difficult category. The overall inter-rater agreement was substantial (Fleiss' κ=0.85), indicating a high level of consistency across raters in difficulty classification.

Considering that completing all 36 questions could be time-consuming and to minimize participant fatigue, the questions were distributed evenly according to difficulty level, and two separate evaluation forms were prepared using Google Documents (Fig. 1). To ensure even distribution, six questions from each of the three difficulty levels determined by the expert panel (EP) were selected, resulting in two evaluation forms, each containing 18 questions.

A total of 155 IDs were included in the study. Participants' personal information was not recorded. The IDs completed the ED rotation in groups over two-month periods throughout the one-year academic calendar, with each group comprising 25–30 IDs. The study was conducted during the final two weeks of each group's two-month ED rotation.

The same questions were subsequently posed to ChatGPT. For the AI component of our study, ChatGPT 4.0 (GPT-4-based, 2025 version) developed by OpenAI was used. Prior to presenting the questions, ChatGPT was provided with contextual information, including the fact that the questions were based on cases presenting to the ED. It was instructed to respond concisely using medical terminology, assuming that the questions were asked by a healthcare professional.

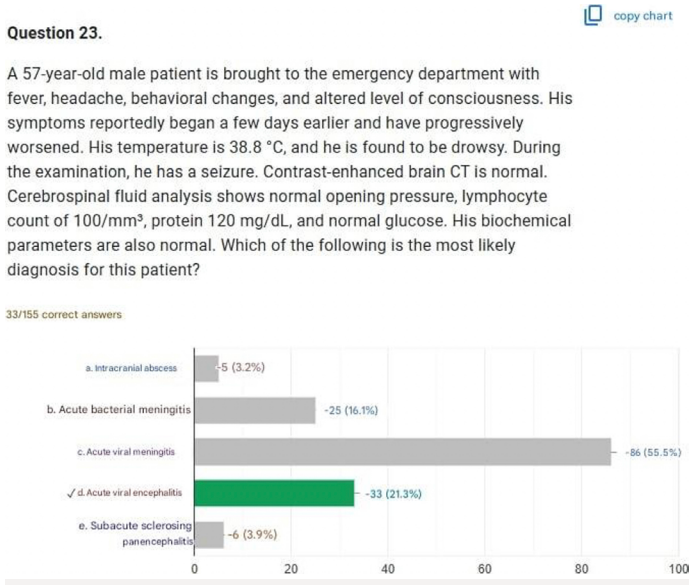

Figure 1. Example of a question completed by intern doctors using Google Documents.

The study was conducted throughout the academic year. Data were collected from IDs during each two-month ED rotation. Initially, a WhatsApp group was created exclusively for IDs on rotation that month, and the link to the Google Documents evaluation form was sent to them. Although the process was conducted online via Google Documents, data collection occurred during the weekly routine ED teaching sessions, with all IDs present and supervised by several proctors, to prevent participants from discussing the questions or consulting written or digital resources. Participants were seated in the classroom in an exam-style arrangement to prevent communication with one another or access to reference materials. Each ID accessed and completed the form using their personal email address, with the form link configured to allow a single view only. Before completing the digitally distributed Google Documents forms, each ID was assigned a unique code, which they entered in the designated field. The same code was used when answering the questions in the second section, ensuring consistency and integrity in data collection. Saving or revisiting the questions was disabled. Subsequently, the questions and correct answers were not shared with any IDs to prevent subsequent groups from becoming aware of the content, thus maintaining optimal blinding.

Responses were evaluated using two different methods. In the first method, all IDs were considered as a single group, and their answers were assessed proportionally. The 50% threshold was used only as a descriptive indicator of how commonly each question was recognized by the intern cohort. This

threshold reflects clinical interpretability rather than statistical classification. In emergency medicine education and Objective Structured Clinical Examination (OSCE)-style.[10,11] competency assessments, a case vignette is typically considered clinically recognizable when at least half of the trainees can identify the correct diagnosis. In the second method, IDs were evaluated individually. Each question was assigned one point (for selecting the correct answer as defined by the gold standard, GS), and scores from each question category were evaluated separately.

Two different approaches were also applied when posing the questions to ChatGPT. Initially, questions were presented in an open-ended format without answer options. The AI was asked to determine the most appropriate diagnosis for each case, and its ability to reach the GS answer was assessed. This section was designated as "GPT-a" (Fig. 2). Subsequently, the questions were presented in a multiple-choice format, similar to how they were posed to IDs, and ChatGPT was asked to select the "most likely diagnosis" from the provided options. Selecting the answer corresponding to the GS was considered a correct response. The results obtained from this phase were designated as "GPT-b" (Fig. 3).

Responses obtained from Google Documents and ChatGPT were collected and compared with the GS answers. The easy, moderate, and difficult question categories were analyzed separately using statistical methods.

Statistical Analysis

Assuming a Type 1 error (α) of 0.05 and a power of 0.95, with an expected joint correct response rate of 0.5 based on expert opinion and a margin of error (d) of 0.2, the sample size required for the study was calculated as 104 intern doctors (IDs) using the confidence interval method. The PASS 11-2025 – Power Analysis and Sample Size Software (v25.0.x) program was used to calculate the sample size. Data analysis was performed using IBM SPSS 25.0 (SPSS Inc., Chicago, IL, USA). Descriptive statistics were presented as mean ± standard deviation and median (min-max) for numerical variables, and as frequency and percentage for categorical variables.

Scores obtained by IDs from easy, moderate, and difficult question categories were compared with scores obtained by ChatGPT in both open-ended and multiple-choice formats using a one-sample t-test. A p-value of <0.05 was considered statistically significant. ChatGPT outputs are deterministic and do not constitute a statistical sample. Since no variance or repeated measurements exist for the AI model, group-to-group comparison tests (e.g., independent-samples t-test) are not statistically appropriate. Thus, intern doctors' mean scores were compared to ChatGPT's fixed score using a one-sample t-test, which evaluates whether a sample mean differs from a fixed, non-random reference value.

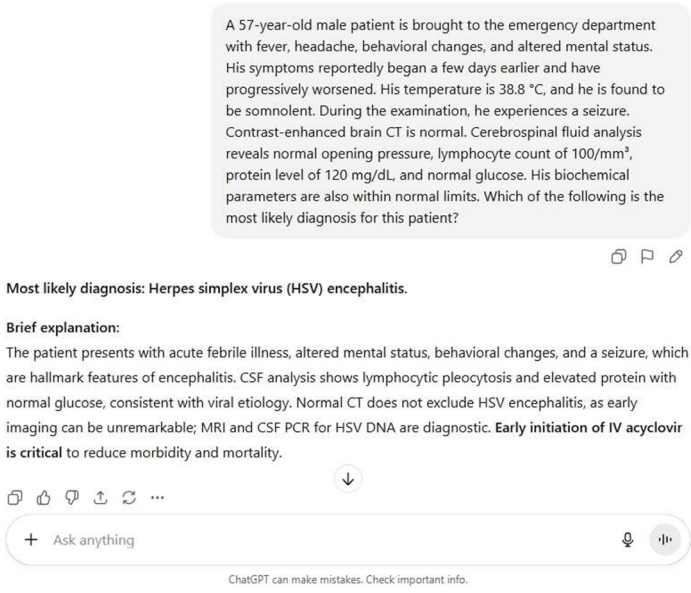

Figure 2. Example of a case presented to Chat-GPT in an open-ended format and its response.

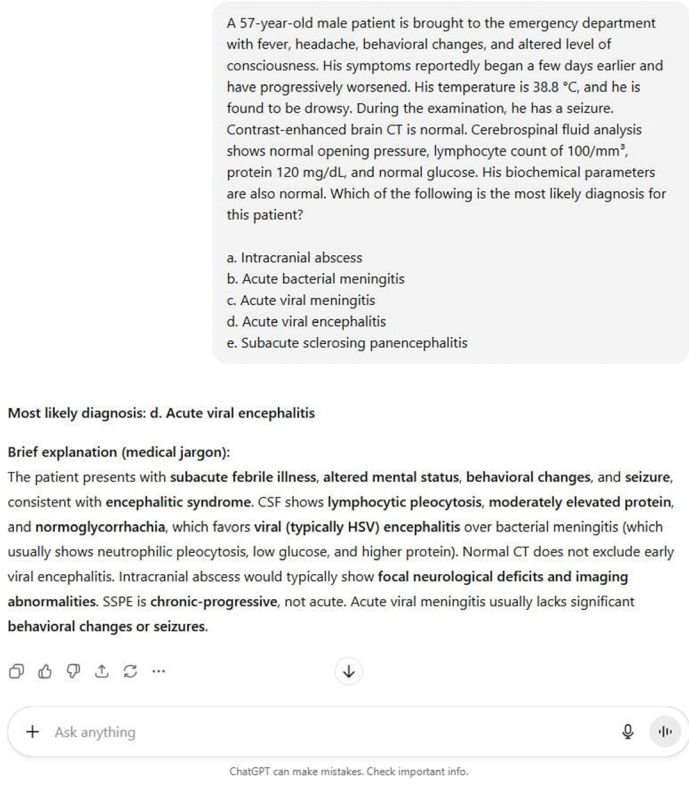

Figure 3. Example of a case presented to Chat-GPT in a multiple-choice format and its response.

Table 1. Comparison of responses given by intern doctors and Chat-GPT based on question difficulty

	Intern doctors n (%)	GPT-a n (%)	GPT-b n (%)
Easy case questions			
True	**11 (30.5)**	8 (22.2)	12 (33.3)
False	1 (2.8)	4 (11.1)	0 (0)
Moderate case questions			
True	**8 (22.2)**	12 (33.3)	12 (33.3)
False	4 (11.1)	0 (0)	0 (0)
Difficult case questions			
True	**2 (5.6)**	8 (22.2)	11 (30.5)
False	10 (27.8)	4 (11.1)	1 (2.8)
Total			
True	**21 (58.3)**	**28 (77.8)**	**35 (97.2)**
False	15 (41.7)	8 (22.2)	1 (2.8)

Table 2. Evaluation of intern doctors' responses based on a percentage system

Easy case questions (%)		Moderate case questions (%)		Difficult case questions (%)	
1	63.9%	13	54.8%	25	**36.8%**
2	93.5%	14	**45.8%**	26	**12.9%**
3	63.9%	15	**45.8%**	27	**26.5%**
4	63.9%	16	51.6%	28	**39.4%**
5	53.5%	17	**43.2%**	29	**36.1%**
6	94.8%	18	60.0%	30	**31.0%**
7	61.9%	19	63.2%	31	**28.4%**
8	58.1%	20	**43.2%**	32	**21.3%**
9	78.7%	21	57.4%	33	**38.1%**
10	93.5%	22	60.0%	34	52.3%
11	78.7%	23	53.5%	35	55.5%
12	**49.0%**	24	70.3%	36	**41.9%**

RESULTS

The responses collected from intern doctors (IDs) via Google Documents were compared with the data obtained from ChatGPT 4.0. It was observed that as question difficulty increased, the accuracy of IDs decreased. In the easy question category, more than 50% of the IDs correctly answered 11 out of 12 questions, whereas this number decreased to 2 out of 12 questions in the difficult category.

When questions were presented to ChatGPT in an open-ended format (GPT-a), it correctly answered all moderate-difficulty questions but provided correct responses for 8 of 12 questions in both the easy and difficult categories. However, in the multiple-choice format (GPT-b), ChatGPT correctly answered all easy and moderate questions, making only one incorrect response in the difficult category.

Overall, the correct response rate for IDs across all questions was 58.3%. In comparison, ChatGPT achieved an accuracy rate of 77.8% in the open-ended format (GPT-a) and 97.2% in the multiple-choice format (GPT-b) (Table 1).

In the easy question category, IDs answered 11 questions correctly, of which ChatGPT correctly answered 8 in the open-ended format and all 11 in the multiple-choice format. The one question answered incorrectly by IDs was correctly identified by ChatGPT when presented in the multiple-choice format. In the moderate category, ChatGPT correctly answered all four questions that IDs had answered incorrectly, in both the open-ended and multiple-choice formats. In the difficult category, ChatGPT correctly answered all 10 questions that IDs failed to answer when presented in the multiple-choice format, and correctly answered 6 of these questions in the open-ended format (Table 1).

The correct response rates for each question among intern doctors (IDs) were analyzed. It was observed that as question difficulty increased, the number of IDs providing correct answers decreased. Correct response rates for difficult questions were significantly lower than those for easy questions. In the easy category, the lowest correct response rate was 49%, while this rate progressively declined with increasing difficulty, with one difficult question being correctly answered by only 12.9% of participants (Table 2).

The data were also analyzed on a question-by-question basis, with one point assigned for each correct answer across the 36 questions. This analysis confirmed that as question difficulty increased, the correct response rate among IDs decreased. A statistically significant, moderate, and negative correlation was observed between IDs' scores and question difficulty ($r=-0.684$; $p<0.001$). Furthermore, the average scores of IDs according to question difficulty were compared with those obtained by ChatGPT in both the open-ended (GPT-a) and multiple-choice (GPT-b) formats, with ChatGPT demonstrating significantly higher performance ($p<0.001$) (Table 3).

DISCUSSION

Our study suggests that ChatGPT-4 could be a valuable tool in the diagnostic process within emergency departments (EDs). It may play a particularly supportive role for newly graduated physicians who lack sufficient clinical experience.

Table 3. Evaluation of intern doctors' responses based on the scoring system and statistical comparison with ChatGPT scores

	Easy case questions	Moderate case questions	Difficult case questions	Total
Intern doctors				
Mean	8.5	6.5	4.2	19.2
(±SD)	1.96	1.81	2.04	4.44
Median	**9**	**6**	**4**	19
Minimum	4	1	0	6
Maximum	12	11	11	30
GPT-a				
Point	8	12	8	28
p value	**0.001**	**<0.001**	**<0.001**	**<0.001**
GPT-b				
Point	12	12	11	35
p value	**<0.001**	**<0.001**	**<0.001**	**<0.001**

SD: Standard deviation.

In a previous study, data from patients admitted from the ED to various departments were retrospectively collected, and both ED residents and ChatGPT-4 were asked, "What could be the diagnosis for these patients?" The results demonstrated that AI provided significantly more accurate responses compared to ED residents ($p<0.05$).[7] In our study, ChatGPT showed statistically significantly higher diagnostic accuracy than intern doctors (IDs) ($p<0.001$). While proportional differences were minimal in easier cases, ChatGPT correctly answered 90% of the questions that IDs had answered incorrectly in more difficult cases.

In another study, ChatGPT-3 was evaluated using 30 case scenarios, presented to both AI and physicians based on limited symptom categories such as abdominal pain, shortness of breath, and vomiting. When ChatGPT-3 was asked to select one diagnosis from ten differential diagnoses in a multiple-choice format, it achieved an accuracy rate of 93.3%. However, when the number of differential diagnoses was reduced to five, its accuracy dropped to 83.3%. When presented with only two answer choices and asked, "What is the most likely diagnosis?", AI correctly answered only 53.3% of the questions.[12]

Our study differs from the aforementioned research. ChatGPT-4 achieved a 97.2% accuracy rate, correctly answering 35 out of 36 multiple-choice questions spanning a wide range of case categories, including varying difficulty levels, traumatic cases, pediatric cases, chronic diseases, and life-threatening clinical conditions. In the previous study, reducing the number of differential diagnoses negatively affected AI's accuracy. In contrast, in our study, ChatGPT-4 achieved a 77.8% accuracy rate even in the open-ended format, without provided answer choices. Considering that IDs achieved a 58.3% success rate in the multiple-choice format, ChatGPT's performance in the open-ended setting can be regarded as highly successful.

The marked improvement of ChatGPT-4 over ChatGPT-3, as observed in our study, underscores the rapid advancements in AI and suggests its potential for even more effective clinical applications in the near future.

In a study by Günay et al.[13] in 2024 on ECG interpretation, ECG findings were converted into text format and presented in a multiple-choice format to experienced ED specialists, cardiologists, and ChatGPT. AI provided more accurate responses than both ED specialists and cardiologists, with statistical significance ($p<0.001$). As the difficulty of the ECG questions increased, the accuracy rates of both physician groups decreased, as did ChatGPT's performance. However, despite this decline, AI remained statistically significantly more successful than both physician groups overall. This study shares similarities with our research in terms of question preparation. In our study, AI was also found to be statistically significantly more successful than intern doctors (IDs) ($p<0.001$). However, whereas no statistically significant difference was observed between cardiologists and ChatGPT in the difficult ECG category in the previous study, we found a statistically significant, moderate, and negative correlation between IDs and ChatGPT as question difficulty increased ($r=-0.684$; $p<0.001$).

Although ChatGPT outperformed IDs overall in our study, it failed to correctly answer 3 questions in the easy category. This may be explained by IDs utilizing their theoretical knowledge to select the correct option from multiple choices, whereas

ChatGPT, applying a more complex reasoning process, sometimes provided multiple possible causes rather than directly stating the correct answer.

As in many areas of life, the potential applications of AI in healthcare have been explored and demonstrated in various experimental studies.[6,14] AI has been shown to be beneficial in clinical settings for tasks such as assigning triage codes, interpreting ECG findings, and analyzing radiological images.[8,9,15] However, studies investigating the use of AI specifically in the diagnostic process of ED cases remain limited.[16] We believe that our study will make a meaningful contribution to the existing literature on this topic.

Although the data on case evaluation in the ED are promising, standardization of ethical considerations, patient consent, and data privacy must be ensured before AI-assisted patient management can be implemented as routine practice.[17]

This study has several important limitations. The most significant limitation is that while one group consisted of scores from 155 intern doctors (IDs), the other group included only a single set of responses obtained from ChatGPT. Since ChatGPT provided a single output per question, its performance does not form a statistical distribution. This required the use of a one-sample comparison approach rather than two-sample methods. Although this is a methodological limitation, it reflects how large language models operate in real clinical use by generating a single diagnostic suggestion.

Another limitation relates to the timing of IDs' ED rotations. Since the study was conducted during the ED rotation, participants joined the study in two-month periods throughout the academic year. Consequently, there was no uniformity regarding the completion of non-ED rotations.

CONCLUSION

It is believed that AI could be a useful tool in the diagnostic phase in EDs. A foundational level of medical knowledge remains essential to correctly input and interpret patient data within AI systems. Therefore, it is essential to recognize that AI should not be considered a standalone diagnostic tool for the general public.

Ethics Committee Approval: Ethics committee approval was obtained from Erzincan Binali Yıldırım University Non-Interventional Clinical Research Ethics Committee (Approval Number: 2024-13/05, Date: 03.10.2024).

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

Conflict of Interest: The authors have no conflicts of interest to declare.

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

Use of AI for Writing Assistance: No use of AI-assisted technologies was declared by the authors.

Author Contributions: Concept – YK, MI, EB, SH, AOE, YB, FMS; Design – YK, MI, EB; Supervision – AOE, YB, FMS; Resource – YK, MI, EB, SH, AOE, YB, FMS; Materials – YK, MI; Data Collection and/or Processing – EB, SH, AOE; Analysis and/or Interpretation – SH; Literature Review – YK, YB, FMS; Writing – YK, MI, EB; Critical Review – SH, AOE, YB, FMS.

Acknowledgment: Authors would like to thank the Proofreading & Editing Office of the Dean for Research at Erciyes University for the copyediting and proofreading service for this manuscript.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Güven G, İnceoğlu F, Özdemir F, Bakır OS. Burnout and professional commitment in medical faculty students in Türkiye: a cross-sectional analysis. Res Dev Med Educ 2023;12:14. [CrossRef]
2. Tortum F, Bayramoğlu A, Tasci HK, Kasali K. Evaluation of the development of emergency response skills of intern doctors after emergency medicine internship: A prospective cross-sectional study. J Res Clin Med 2023;11:13. [CrossRef]
3. Turan S, Hayme S. Effect of professional commitment on the career plans of intern doctors. Pak J Med Sci 2025;41(2):525-30. [CrossRef]
4. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt Accessed March 31, 2026.
5. Caldarini G, Jaf S, McGarry K. A Literature Survey of Recent Advances in Chatbots. Information 2022;13(1):41. [CrossRef]
6. Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT. Am J Emerg Med 2024;78:170-5. [CrossRef]
7. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. J Med Internet Res 2024;26:e56110. [CrossRef]
8. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv [Preprint] 2023:2023.02.02.23285399. Update in: J Am Coll Radiol 2023;20(10):990-7. [CrossRef]
9. Zaboli A, Brigo F, Ziller M, Massar M, Parodi M, Magnarelli G, et al. Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department. Am J Emerg Med 2025;88:7-11. [CrossRef]

10. Muralidharan J, Lewin S, Nithyanandam S, Stephen J, Joseph M, Ta L, et al. Feasibility and utility of objective structured clinical examination (OSCE) as an end-of-internship assessment of MBBS training in a medical college in Southern India. BMC Med Educ 2025;25(1):1608. [CrossRef]
11. Lee M, Hernandez E, Brook R, Ha E, Harris C, Plesa M, et al. Competency-based Standard Setting for a High-stakes Objective Structured Clinical Examination (OSCE): Validity Evidence. MedEdPublish (2016) 2018;7:200. [CrossRef]
12. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health 2023;20(4):3378. [CrossRef]
13. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024;80:51-60. [CrossRef]
14. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): A preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023;23(3):156-61. [CrossRef]
15. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med 2024;79:44-7. [CrossRef]
16. Farahmand S, Shabestari O, Pakrah M, Hossein-Nejad H, Arbab M, Bagheri-Hariri S. Artificial Intelligence-Based Triage for Patients with Acute Abdominal Pain in Emergency Department; a Diagnostic Accuracy Study. Adv J Emerg Med 2017;1(1):e5.
17. Farangi MR, Nejadghanbar H, Hu G. Use of generative AI in research: ethical considerations and emotional experiences. Ethics & Behavior 2025;35(7):527-43. [CrossRef]