EURASIAN JOURNAL OF EMERGENCY MEDICINE
Original Article
DOI: 10.4274/eajem.galenos.2025.60320
Eurasian J Emerg Med. 2025;24(3): 201-7

Large Language Models in Generating Differential Diagnoses in the Emergency Department: A Comparative Study of ChatGPT, Copilot, and Emergency Physician

Banu Arslan[1], Büşra Erdem[2], Merve Osoydan Satıcı[2], Çağatay Nuhoğlu[2], Hasan Yasin Soylu[1]

[1]University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye
[2]University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye

Abstract

Aim: Accurate diagnosis in emergency departments relies heavily on clinical decision-making, yet cognitive errors contribute to a significant proportion of diagnostic mistakes. Since their launch, Generative Pre-trained Transformer-4 (GPT-4) based large language models (LLMs) have been reshaping healthcare, offering improvements in diagnostic accuracy, treatment planning, and patient care. This study evaluates the performance of these tools in generating primary and differential diagnoses compared to an experienced emergency medicine (EM) physician.

Materials and Methods: We conducted a retrospective cross-sectional study using 468 real-world clinical vignettes from non-trauma adult patients. GPT-4-based Chat Generative Pre-trained Transformer (ChatGPT) and Copilot were tasked with generating five differential diagnoses for each vignette. Their accuracy was compared to the diagnoses provided by EM physicians, using discharge diagnoses as the reference. Statistical analysis included descriptive statistics and Cohen's kappa to assess agreement.

Results: ChatGPT and Copilot demonstrated high accuracy, with correct diagnoses in the top three positions in 91.9% and 90.2% of cases, respectively, compared to 93.2% for the EM physician. Moderate agreement between the artificial intelligence (AI) tools and the EM physician was observed (kappa: 0.476 for ChatGPT and 0.414 for Copilot).

Conclusion: LLM-based generative AI tools show promise as clinical decision support systems, enhancing diagnostic accuracy and assisting less-experienced clinicians. However, they should complement, not replace, human expertise in emergency settings.

Keywords: ChatGPT, Copilot, LLMs, generative artificial intelligence, differential diagnosis, emergency

Introduction

The diagnostic process relies on four essential components: medical history, physical examination, differential diagnoses, and diagnostic tests. Experienced clinicians have historically been able to diagnose accurately in 70%, 90% of cases by taking a thorough and detailed medical history (1,2). This demands extensive medical knowledge, sharp observational skills, and rigorous logical reasoning. The same precision is required when formulating a comprehensive list of differential diagnoses, as any error in judgment can derail the entire diagnostic process, and lead to a delayed or missed diagnosis with potentially serious consequences for patient outcomes. Unfortunately, approximately 7.4 million (5.7%) emergency department (ED) visits in the USA involve at least one diagnostic error annually causing 371,000 (0.3%) serious harms (3). A total of 89% of these errors are attributed to failures in clinical decision-making or judgment, regardless of the underlying disease. Cognitive errors causing delayed or missed diagnosis that are related to human factors such as clinical expertise, inadequate knowledge, or critical thinking have been reported in several studies (4-7).

A potential solution for mitigating cognitive errors in the diagnostic process is employing artificial intelligence (AI) technologies. Brown et al. (8) reported that improved clinical

Corresponding Author: Banu Arslan MD, University of Health Sciences Türkiye, Başakşehir Çam and Sakura City, Clinic of Emergency Medicine, İstanbul, Türkiye
E-mail: dr.banuarslan@gmail.com **ORCID ID:** orcid.org/0000-0003-0982-5351

Cite this article as: Arslan B, Erdem B, Osoydan Satıcı M, Nuhoğlu Ç, Soylu HY. Large language models in generating differential diagnoses in the emergency department; a comparative study of ChatGPT, Copilot, and emergency physician. Eurasian J Emerg Med. 2025;24(3): 201-7.

Received: 11.02.2025
Accepted: 01.05.2025
Epub: 07.07.2025
Published: 10.09.2025

decision-making and reduced risk of patient harm could be achieved in emergency patients by employing AI technologies, including AI-based symptom checkers, natural language processing tools to generate differential diagnoses, and real-time electrocardiography and X-ray interpretation tools. Similarly, Harada et al. (9) and Schwitzguebel et al. (10) reported that less-experienced physicians could enhance their diagnostic accuracy by using AI technologies in history taking and generating differential diagnoses.

With the launch of Chat Generative Pre-trained Transformer (ChatGPT) (OpenAI, San Francisco, CA) in November 2022 and Copilot (formerly Bing AI) (Microsoft, Redmond, WA) in February 2023, AI technologies became easier for the general public to use and more accessible. Their ability to comprehend inputs and generate human-like, fluent text outputs to queries, quickly drew attention across various fields, including healthcare. The latest versions of these models were built on the Generative Pre-trained Transformer-4 (GPT-4) architecture, which utilizes a transformer-based neural network to predict the next token in a document (11). GPT-4's advanced capabilities make it particularly useful in healthcare, where it can assist health professionals with tasks such as medical diagnosis and generating differential diagnosis lists. A user can input a patient's clinical manifestations into these large language model (LLM)-based generative AI tools, allowing the models to analyze medical text data and suggest potential differential diagnoses. In this research, we aimed to determine the diagnostic accuracy of these novel AI tools using real-life clinical vignettes and to have them list differential diagnoses. We believe that this approach can serve as a clinical decision support system (CDSS) tool, providing valuable assistance to less experienced health professionals in their practice in the future.

Materials and Methods

Study Setting and Data Collection

The present study was designed as a single-center, retrospective and cross-sectional study and conducted with 468 real-world clinical vignettes. The data for the 468 patients were obtained from the previous research titled "evaluating LLM-based generative AI tools in a five-level emergency triage system: a comparative study of ChatGPT Plus, Copilot Pro, and triage nurses". That study was conducted in an ED of a large urban academic hospital over a one-week period between December 11 and December 18, 2023. Adult patients were enrolled during random 24-hour intervals. Exclusions included minors, trauma cases, and incomplete data. In the triage area, nurses assessed patients while the emergency physician observed them and then documented standardized clinical vignettes.

For our research, Institutional Review Board approval was obtained University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital Health Practice and Research Clinical Research Ethics Committee (decision number: 4640, date: 03.12.2024).

Outcome and Procedure

The primary outcome of the study was determining the accuracy of LLM-based AI applications in generating primary and differential diagnoses for non-trauma adult patients presenting to the ED. To do so, we simultaneously introduced each clinical vignette to the GPT-4 based generative AI tools and an experienced emergency medicine (EM) physician, by asking "Can you list 5 possible diagnoses, ordered from most likely to less likely based on the presented information above?" The first diagnosis was accepted as the primary diagnosis and the others as differential diagnoses. These diagnosis lists were then evaluated by an academic EM physician and benchmarked against actual discharge diagnoses. The accuracy of LLM-based generative AI tools and EM specialists in identifying the correct diagnosis was investigated by determining the position of the final diagnosis in each list.

LLMs Based-generative AI tools

Two LLM-based generative AI tools were used in this study; ChatGPT Plus and Copilot Pro. ChatGPT Plus was accessed via OpenAI's GPT-4 interface (version dated: 11.12.2023), and Copilot Pro was used via Microsoft Copilot with GPT-4 integration as of the same date. These tools can be considered advanced AI systems that can create high-quality content, such as text, audio, code, images, and videos, based on the data they were trained on. They can analyze raw data and identify underlying patterns and structures, allowing them to generate the most statistically likely outputs in response to specific prompts. Both models ran on Open AI's GPT-4 and performed various tasks. However, its infrastructure is not publicly disclosed.

Statistical Analysis

Statistical analyses were performed using SPSS software, version 28.0. Descriptive statistics, including means, standard deviations, medians, minimum and maximum values, frequencies, and percentages, were used to summarize and characterize the data. The distribution of variables was checked with the Kolmogorov-Smirnov test. The Friedman test and the Wilcoxon test were used for the repeated measurement analysis. The chi-square test was used for comparison of qualitative data. Cohen's kappa was used to assess the level of agreement between LLMs and EM physicians.

Results

A total of 468 clinical vignettes were included in the study. Patients ranged in age from 18 to 100 years, with a mean age of 46±19.8 years and a median age of 44.5 years. The gender distribution was balanced, with 54.7% female (n=256) and 45.3% male (n=212). The majority presented with ear, nose, and throat (ENT) symptoms (15.4%, n=72), followed by respiratory (15.2%, n=71) and cardiovascular complaints (13.9%, n=65). In terms of comorbidities, 63.7% (n=298) had no comorbidities, while 36.3% (n=170) had at least one. The most common comorbidities were hypertension (16.2%, n=76), diabetes mellitus (9.6%, n=45), and coronary artery disease (6.6%, n=31). Most patients (91.2%, n=427) were discharged; 4.5% required gastrointestinal (GI) hospitalization, 3.2% were admitted to the intensive care unit (ICU), 0.9% refused care, and 0.2% died. Demographic data were presented in Table 1.

The accuracy of the EM physician, ChatGPT, and Copilot in predicting the correct diagnosis within the top five listed diagnoses was 98.5%, 97.4%, and 94.9%, and the top three listed diagnoses was 93.2%, 91.9%, and 90.2%, respectively. The EM physician identified the definitive diagnosis as the first choice in 379 cases, with an accuracy rate of 81.0%, compared to ChatGPT's 351 cases (75%) and Copilot's 345 cases (73.7%) (Table 2).

The EM physician failed to include the final diagnosis within the top five differential diagnoses in 7 cases, compared to 12 missed cases by ChatGPT and 24 by Copilot. Among clinical systems, the EM physician missed 1 ENT case, 2 GI cases, 1 musculoskeletal case, 1 genitourinary (GU) case, and 2 hematologic cases. ChatGPT missed 2 ENT, 1 cardiovascular, 1 GI, 3 musculoskeletal, 1 GU, 3 hematologic, and 1 psychiatric case. Copilot missed 4 ENT, 2 cardiovascular, 4 GI, 9 musculoskeletal, 2 hematologic, and 3 psychiatric cases. In terms of patient outcomes, the EM physician missed 5 discharged cases and 2 ward admissions, with no missed ICU or surgical/intervention cases. ChatGPT missed 9 discharged cases, 2 ward admissions, and 1 ICU admission. Copilot missed 17 discharged cases, 4 ward admissions, 3 cases requiring procedural or surgical intervention, but no ICU admissions (Table 3).

ChatGPT and the EM physician agreed on the rank of the 375 cases; 337 were identified as the first choice. Similarly, Copilot and the EM physician agreed on 361 cases, with 329 being identified as the first choice. There was a moderate agreement between all raters with Cohen's kappa values for the EM physician versus ChatGPT and for the EM physician versus Copilot being 0.476 and 0.414, respectively (p=0.000) (Table 4).

Table 1. Patient demographics, chief complaints, chronic medical conditions, and patient outcomes

		Minimum-Maximum			Median	Mean ± SD/n-%		
Age		18.0	-	100.0	44.5	46.2	±	19.8
Gender	Female					256		54.7%
	Male					212		45.3%
Chief complaints	ENT-mouth, throat, neck					72		15.4%
	Respiratory					71		15.2%
	Cardiovascular					65		13.9%
	Neurological					61		13.0%
	GI					58		12.4%
	Musculoskeletal (limp/joint pain, neck pain)					35		7.5%
	Ophthalmology					23		4.9%
	GU					20		4.3%
	Mental health					18		3.8%
	ENT- nose/ear					18		3.8%
	Skin					17		3.6%
	Fever					8		1.7%
	Poisoning					2		0.4%

Table 1. Patient demographics, chief complaints, chronic medical conditions, and patient outcomes				
		Mean ± SD/n-%		
Comorbidities	None	298		63.7%
	At least 1	170		36.3%
	At least 2	77		16.5%
	3 or more	41		8.8%
	HT	76		16.2%
	DM	45		9.6%
	CAD	31		6.6%
	Malignancy	22		4.7%
	COPD	14		3.0%
	CHF	13		2.8%
	CVD	13		2.8%
	CRF	12		2.6%
	Asthma	10		2.1%
	Migraine	9		1.9%
	Epilepsy	7		1.5%
	Others	30		6.4%
Outcome	Discharge	427		91.2%
	Hospitalization	21		4.5%
	ICU	15		3.2%
	Mortality	1		0.2%
	Refusal of care	4		0.9%

ENT: Ear, nose, and throat, SOB: Shortness of breath, HT: Hypertension, GI: Gastrointestinal, GU: Genitourinary, DM: Diabetes mellitus, CAD: Coronary artery disease, COPD: Chronic obstructive pulmonary disease, CHF: Congestive heart failure, CVD: Cerebrovascular disease, CRF: Chronic renal failure, ICU: Intensive care unit, SD: Standard deviation

Table 2. Comparison of diagnostic rankings among emergency medicine physician, ChatGPT, and Copilot							
		EM physician		ChatGPT		Copilot	
		n	%	n	%	n	%
Diagnostic rank	1	379	81.0	351	75	345	73.7
	2	37	7.9	51	10.9	49	10.5
	3	20	4.3	28	6.0	28	6.0
	4	15	3.2	14	3.0	11	2.4
	5	10	2.1	12	2.6	11	2.4
	≥6	7	1.5	12	2.6	24	5.1

EM: Emergency medicine, ChatGPT: Chat Generative Pre-trained Transforme. This table compares the performance of an emergency medicine physician, ChatGPT, and Copilot in ranking the correct diagnosis among a list of potential diagnoses, with the rankings shown from 1 (most accurate) to 5 (least accurate)

Table 3. Number of missed diagnoses across clinical systems and patient outcomes, as identified by the EM physician, ChatGPT, and Copilot. A missed case was defined as the failure to include the final diagnosis within the top five differential diagnoses generated for each patient vignette

Systems	EM physician	ChatGPT	Copilot
ENT	1	2	4
Cardiovascular system	-	1	2
GI system	2	1	4
Musculoskeletal system	1	3	9
GU system	1	1	
Hematology	2	3	2
Mental health	-	1	3
Patient outcomes			
Discharged	5	9	17
Procedural/surgical intervention	-	-	3
Admitted to a ward	2	2	4
ICU	-	1	-

ENT: Ear, nose, and throat, EM: Emergency medicine, GI: Gastrointestinal, GU: Genitourinary, ICU: Intensive care unit, ChatGPT: Chat Generative Pre-trained Transforme

Table 4. Kappa accuracy test between EM physician and LLMs

		Emergency medicine physician						Accuracy	p value
Diagnostic rank		**1**	**2**	**3**	**4**	**5**	**≥6**		
ChatGPT	1	**337**	8	2	4	0	0	80.1%	Kappa: 0.476 p=0.000
	2	22	**20**	3	3	2	1		
	3	9	6	**7**	3	3	0		
	4	5	3	2	**3**	1	0		
	5	5	0	1	2	**3**	1		
	≥6	1	0	5	0	1	**5**		
Copilot	1	**329**	12	2	2	0	0	77.1%	Kappa: 0.414 p=0.000
	2	21	**17**	6	3	2	0		
	3	14	5	**5**	3	1	0		
	4	5	1	2	**2**	1	0		
	5	5	1	0	2	**2**	1		
	≥6	5	1	5	3	4	**6**		

EM: Emergency medicine, LLMs: Large language models, ChatGPT: Chat Generative Pre-trained Transforme

Discussion

This study provides an in-depth analysis of the diagnostic performance of the GPT-4-based ChatGPT and Copilot using real-world patient data. Our findings indicate that both tools exhibit a high level of accuracy in predicting the correct diagnosis based on patients' clinical history and vital parameters. This supports the growing body of evidence suggesting that AI-driven models can enhance diagnostic accuracy in clinical settings.

Previous studies have also reported promising results regarding AI's ability to assist in diagnosis. For example, Levine et al. (12) evaluated a GPT-3-based AI model (an earlier version of ChatGPT) using 48 clinical vignettes and found that the model identified the correct diagnosis in the top three differential diagnoses with an accuracy of 88%. Similarly, another study using 30 clinical vignettes reported an accuracy rate of 83.3% in generating the correct diagnosis within a list of five possible differentials (13). More recent research conducted by Hirosawa et al. (14) has demonstrated the superiority of newer AI models, including ChatGPT-3.5 and ChatGPT-4.0. In that study, ChatGPT-3.5 achieved a 65% accuracy rate in listing the correct diagnosis within the top five differential diagnoses, while ChatGPT-4.0 improved this rate to 81%. In line with these previous studies, our research highlights the notable diagnostic performance of GPT-based

LLMs in an emergency setting. Both tools demonstrated a high degree of accuracy when listing differential diagnoses, achieving rates above 90% for the top three predictions. This improvement in performance over earlier versions of these models can be attributed to advances in model architecture, increased training data, and enhanced fine-tuning.

Notably, some of the missed diagnoses we identified involved potentially serious conditions, such as cardiovascular or hematologic disorders, which could delay critical interventions. Specifically, Copilot failed to include the correct diagnosis in three cases, that ultimately required procedural or surgical treatment, two of which were acute limb ischemia. Similarly, ChatGPT missed a diagnosis in a patient later admitted to the ICU, indicating that diagnostic oversights by AI tools could have meaningful clinical consequences. These findings highlight the importance of cautious integration of generative AI in emergency decision-making and underscore the continued need for clinical oversight.

Another notable finding in our study is the moderate agreement between the LLM-based tools and the EM physician, as reflected by the Cohen's kappa values. While accuracy was a key measure, we also evaluated this parameter to explore the potential of AI tools as supportive aids in clinical decision-making. Continuous feedback is essential for improving diagnostic accuracy (15) and enabling the smooth integration of AI into clinical workflows. Although research in this area is limited, a recent study involving 392 case descriptions reported similar results, with moderate agreement between GPT-4 and physicians (Cohen's kappa= 0.47) (0.39-0.56) in generating differential diagnoses (16). This level of agreement between EM physicians and LLM-based AI highlights their potential value, especially for supporting less-experienced clinicians in decision-making.

In light of these considerations, we propose that LLM-based generative AI tools such as ChatGPT and Copilot should be integrated into healthcare as CDSS, particularly in high-volume settings like the ED, where time constraints, high levels of stress and diagnostic complexity are prevalent. AI tools can assist in providing rapid, evidence-based suggestions, ensuring that fewer diagnoses are overlooked. This approach can help mitigate the cognitive biases that often contribute to diagnostic errors in emergency care.

Study Limitations

This study has three main limitations. First, data were collected over one week in December, a period with high rates of upper respiratory tract infections. This short duration may not fully capture seasonal variations in disease presentations and patient demographics, potentially affecting diagnostic accuracy and the study's generalizability. Second, this study used the discharge diagnosis as the final diagnosis. While discharge diagnosis serves as a reasonable benchmark, it may not always reflect the most accurate final diagnosis due to misdiagnosis. Third, our study was conducted in a controlled environment rather than real emergency settings. In real-time clinical practice, cognitive ability is influenced by multiple dynamic factors such as time constraints, high patient volume, physician workload, stress, and cognitive fatigue. These factors may affect clinical decision-making, ranking, and variety of potential diagnoses. Given these considerations, we recommend that future studies be conducted in real-time and real-world emergency settings.

Conclusion

Our study demonstrates the significant potential of LLM-based generative AI tools like ChatGPT and Copilot to assist clinicians in diagnostic reasoning. These tools should be viewed as complementary aids rather than replacements for human expertise. To ensure safe and effective integration into clinical practice, their implementation must be accompanied by continuous evaluation. While our findings are promising, future multicenter studies are needed to enhance generalizability and validate performance across diverse clinical settings.

Ethics

Ethics Committee Approval: Institutional Review Board approval was obtained University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital Health Practice and Research Clinical Research Ethics Committee (decision number: 4640, date: 03.12.2024).

Informed Consent: The present study was designed as a single-center, retrospective and cross-sectional study and conducted with 468 real-world clinical vignettes.

Footnotes

Authorship Contributions

Concept: B.E., Design: B.A., M.O.S., Ç.N., H.Y.S., Data Collection or Processing: B.A., B.E., Analysis or Interpretation: B.A., B.E., M.O.S., Ç.N., H.Y.S., Literature Search: B.A., B.E., Writing: H.Y.S.

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. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med. 1992;156:163-5.
2. Gruppen LD, Woolliscroft JO, Wolf FM. The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. Res Med Educ. 1988;27:242-7.
3. Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, et al. Diagnostic errors in the emergency department: a systematic review [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2022.
4. Hussain F, Cooper A, Carson-Stevens A, Donaldson L, Hibbert P, Hughes T, et al. Diagnostic error in the emergency department: learning from national patient safety incident report analysis. BMC Emerg Med. 2019;19:77.
5. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med. 2007;49:196-205.
6. Norman G. Building on experience--the development of clinical reasoning. N Engl J Med. 2006;355:2251-2.
7. Committee on diagnostic error in health care; board on health care services; institute of medicine; the national academies of sciences, engineering, and medicine. Improving diagnosis in health care. Balogh EP, Miller BT, et al. Washington (DC): National Academies Press (US); 2015.
8. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking bias: the role of artificial intelligence in improving clinical decision-making. Cureus. 2023;15:e36415.
9. Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. Int J Environ Res Public Health. 2021;18:2086.
10. Schwitzguebel AJ, Jeckelmann C, Gavinio R, Levallois C, Benaïm C, Spechbach H. Differential diagnosis assessment in ambulatory care with an automated medical history-taking device: pilot randomized controlled trial. JMIR Med Inform. 2019;7:e14044.
11. OpenAI. GPT-4 Technical Report. Preprint at arXiv https://doi.org/10.48550/arXiv.2303.08774 (2023).
12. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health. 2024;6:e555-61.
13. 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;20:3378.
14. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. 2023;11:e48808.
15. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. BMJ. 2022;376:e068044.
16. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. 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.