Thorac Res Pract. 2026;27(1):38-46 **DOI:** 10.4274/ThoracResPract.2025.2025-6-3

Original Article

AI in Patient Care: Evaluating Large Language Model Performance Against Evidence-Based Guidelines for Pulmonary Embolism

Ömer F. Karakoyun[1], Halil E. Koyuncuoğlu[1], Ömer H. Sağnıç[2], Mehmed E. Özdemir[3], Yalçın Gölcük[4], Birdal Yıldırım[4]

[1]Clinic of Emergency Medicine, Muğla Training and Research Hospital, Muğla, Türkiye
[2]Clinic of Emergency Medicine, Isparta City Hospital, Isparta, Türkiye
[3]Department of Artificial Intelligence, Muğla Sıtkı Koçman University, Graduate School of Natural and Applied Sciences, Muğla, Türkiye
[4]Department of Emergency Medicine, Muğla Sıtkı Koçman University Faculty of Medicine, Muğla, Türkiye

Cite this article as: Karakoyun ÖF, Koyuncuoğlu HE, Sağnıç ÖH, Özdemir ME, Gölcük Y, Yıldırım B. AI in patient care: evaluating large language model performance against evidence-based guidelines for pulmonary embolism. *Thorac Res Pract.* 2026;27(1):38-46

ABSTRACT

OBJECTIVE: Artificial intelligence (AI)-driven large language models (LLMs) are increasingly used in patient education; however, their ability to interpret and apply clinical guidelines within real-world physician workflows remains uncertain. Pulmonary embolism (PE), with its well-established diagnostic and management protocols, provides a suitable model for evaluating these systems. This study assessed the performance of four widely used AI-driven LLMs—ChatGPT-4o, DeepSeek-V2, Gemini, and Grok—in applying the 2019 European Society of Cardiology guidelines for PE. The focus was on evaluating clinical accuracy, adherence to guidelines, and response consistency.

MATERIAL AND METHODS: Ten open-ended questions based on a simulated PE case were created, covering diagnosis, risk stratification, treatment, and follow-up. Guideline-based reference answers were used for scoring. LLMs were tested under identical conditions, and the responses were anonymized and scored by two emergency physicians using a 10-point scale. Inter-rater reliability was measured using the intraclass correlation coefficient (ICC), and group comparisons were made using Kruskal-Wallis tests.

RESULTS: ChatGPT-4o scored highest (76), followed by Gemini (73.75), Grok (71.25), and DeepSeek-V2 (65). No significant difference was found in total scores ($P = 0.390$). Performance varied by category; ChatGPT-4o excelled in follow-up, while DeepSeek-V2 performed best in diagnostics. Expert reviewers noted ChatGPT-4o's structured responses and Grok's practicality, but highlighted limitations such as insufficient personalization and guideline gaps. Inter-rater agreement was excellent (ICC: 0.986).

CONCLUSION: AI-driven LLMs show promise in supporting PE management, though none consistently excel in all domains. Further development is needed to enhance clinical integration and guideline compliance.

KEYWORDS: Pulmonary embolism, artificial intelligence, clinical decision support

Received: 23.06.2025 ***Revision Requested:*** 15.11.2025 ***Last Revision Received:*** 18.11.2025

Accepted: 22.12.2025 ***Epub:*** 20.01.2026 ***Publication Date:*** 30.01.2026

INTRODUCTION

Technological advancements have significantly transformed patient care by improving diagnostic accuracy, optimizing treatment planning, and enhancing overall healthcare efficiency. In recent years, artificial intelligence (AI)-driven large language models (LLMs) have emerged as potential clinical decision-support tools capable of assisting physicians in real time during routine patient management, particularly by synthesizing guideline-based recommendations for complex clinical scenarios. LLMs, designed to comprehend, process, and generate human language, are AI-driven systems trained on predefined datasets to respond efficiently to a wide range of queries and to retrieve accurate information from the internet, using advanced natural language processing (NLP) models. Current LLM technologies are capable of extracting evidence-

Corresponding author: Prof. Birdal Yıldırım, MD, e-mail: birdalgul@gmail.com

based information and presenting it in a natural conversational format.[1] Consequently, AI-driven LLMs have the potential to serve as valuable point-of-care reference tools by delivering clinical information in a clear, interactive, and context-specific manner. Given their capacity to process extensive medical literature and rapidly evolving clinical protocols, LLMs hold promise as point-of-care reference tools that may support physicians' diagnostic reasoning, risk stratification, and management planning.

The increasing body of literature on the applicability of AI in the medical field highlights its expanding role in healthcare. Studies evaluating the responses of AI-driven LLMs with NLP capabilities to patient inquiries have demonstrated that the accuracy and reliability of these responses often meet medically acceptable standards.[2-4] These findings suggest that AI-driven LLMs hold promise as potential medical advisors, offering reliable health-related information and assisting patients in understanding their conditions. While further validation and regulatory oversight are necessary, the growing evidence supports the potential integration of AI-driven systems into patient education and preliminary medical consultations.

Given the rapid advancements in AI and NLP technologies, LLMs have the potential to bridge this gap by providing physicians with quick and accurate access to updated clinical guidelines. By processing vast amounts of medical literature in real time, AI-driven systems can assist healthcare professionals in retrieving guideline-based recommendations relevant to specific clinical scenarios. Moreover, with the integration of LLMs that incorporate deep learning methods tailored for healthcare professionals, NLP-powered systems have significant potential to analyze and interpret clinical guidelines with greater precision.[5] Additionally, these systems can help synthesize complex information, highlight key updates, and present context-specific recommendations, thereby improving adherence to evidence-based practices. However, the reliability, accuracy, and clinical applicability of AI-generated recommendations remain critical areas that require further investigation and validation.

This study aims to assess the performance of four widely used AI-driven LLMs in managing simulated patient cases across various clinical scenarios. Specifically, we will examine their ability to apply the 2019 European Society of Cardiology (ESC) guidelines for the management of pulmonary embolism (PE) and will evaluate the accuracy of their responses and their adherence to established recommendations.[6] By comparing their outputs with evidence-based standards, we seek to determine their potential role as clinical decision-support tools. Assessing how LLMs handle guideline-driven decision steps within PE management provides meaningful insight into their potential integration into physician workflows.

Main Points

- This study evaluated the clinical accuracy and guideline adherence of four widely used artificial intelligence (AI) chatbots in managing a simulated case of acute pulmonary embolism based on the 2019 European Society of Cardiology guidelines.
- ChatGPT-4o demonstrated the highest overall score, while DeepSeek-V2 had the lowest performance, particularly in risk stratification and treatment planning.
- Despite individual strengths, none of the AI models consistently excelled across all phases of diagnosis, treatment, and follow-up, highlighting variability in clinical applicability.
- Expert reviewers found notable differences in the models' ability to provide structured, evidence-based, and patient-specific responses.

MATERIAL AND METHODS

Study Design and Setting

This study was designed as a comparative evaluation of AI-driven LLMs for patient management, specifically assessing their adherence to the 2019 ESC guidelines for the diagnosis and management of PE. Since the study does not involve human participants, patient data, or personally identifiable information, ethical approval was deemed unnecessary. All analyses were conducted using AI-generated responses to predefined clinical scenarios, ensuring a standardized and controlled assessment environment. This research adheres to ethical considerations relevant to AI-based studies in medicine and aligns with the principles outlined in the Declaration of Helsinki.

Development of Clinical Scenarios and Questions

A simulated clinical case was developed to reflect real-world patient presentations of PE. Based on this scenario, ten open-ended questions were formulated, each addressing a key aspect of PE diagnosis, risk stratification, and management as outlined in the 2019 ESC guidelines. The questions were designed to assess the ability of LLMs to generate evidence-based responses with an appropriate level of clinical accuracy.

To ensure standardization, guideline-based reference answers were pre-constructed for each question using a 10-point scoring system. These answers incorporated predefined key points, enabling an objective and structured evaluation of LLM-generated responses. The complete list of questions and their corresponding reference answers is provided in Table 1.

Selection of Artificial Intelligence-Driven Large Language Models

To ensure a broad and representative comparison, four widely used and advanced AI-driven LLMs were selected based on their global popularity, technological diversity, and NLP capabilities. The selected LLMs were:

ChatGPT-4o (OpenAI, USA): Chosen for its strong language understanding, widespread use, frequent updates, and prominence in research.

DeepSeek-V2 (DeepSeek AI, China) was selected to represent Eastern AI development owing to its rising popularity and notable performance gains.

Gemini (PaLM 2 Pro/Ultra) (Google, USA): Known for advanced search integration and robust data processing, backed by Google's AI expertise.

Table 1. Acute pulmonary embolism case-based assessment and scoring criteria

Questions	**Predefined answer key**
Case scenario A 51-year-old female patient was transferred to our emergency department from another healthcare facility with a preliminary diagnosis of syncope of unknown etiology. Her husband reported that, after a 20-hour flight from the United States, she initially experienced mild shortness of breath at the airport, which was followed by a complete loss of consciousness. Since he was holding her at that moment, no fall or trauma occurred. The loss of consciousness lasted 3–4 minutes, without excessive salivation or secretion, tongue biting, or uncontrolled convulsive movements of the limbs. The husband reported that his wife had no history of epilepsy or previous episodes of unconsciousness.	
Question 1: Based on the patient's presenting complaint, what are your differential diagnoses?	If the most probable diagnosis is pulmonary embolism (PE) → **5 points** If any of the following additional possible diagnoses are included, add **1 point** for each: Vasovagal syncope (21%) → **1 point** Cardiac syncope (10%) → **1 point** Orthostatic syncope (9%) → **1 point** Seizure (5%) → **1 point** Neurological causes (4.1%) → **1 point** **Maximum possible score: 10 points**
Case scenario continued The patient's family history is notable for venous thromboembolism in the mother. In terms of medical history, the patient has been using oral contraceptives for the past six years. On physical examination, the patient appeared to be in moderate condition. She is confused, partially oriented, and partially cooperative. Her vital signs indicate hemodynamic instability, with a blood pressure of 85/50 mmHg, a heart rate of 128 bpm, a respiratory rate of 30 breaths per minute, and an oxygen saturation of 88% on room air. Further examination reveals jugular venous distension and significant bilateral lower extremity edema. Additionally, there is notable tenderness in the right calf. Electrocardiographic evaluation demonstrates sinus tachycardia and an incomplete right bundle branch block.	
Question 2: Based on the patient's history, family history, and physical examination findings, what are the most likely preliminary diagnoses? Explain.	If the primary preliminary diagnosis is PE → **4 points** If the response includes relevant risk factors (e.g., prolonged immobilization, long-haul flight, oral contraceptive use) → **2 points** If the response includes a calculated Wells score → **4 points** **Maximum possible score: 10 points**
Question 3: Which laboratory tests would you order to confirm your diagnosis? Additionally, which tests would help assess the patient's prognosis and risk of mortality? Provide a rationale for your choices.	If the answer includes B-type natriuretic peptide (for right ventricular dysfunction and prognosis in PE) → **3 points** If the answer includes Troponin (for myocardial injury and right ventricular strain assessment) → **3 points** If the answer includes arterial blood gas (to assess hypoxemia, hypercapnia, respiratory alkalosis, and elevated lactate as a marker of tissue hypoxia and hemodynamic instability in PE) → **2 points** If the answer includes any of the following additional tests, add 0.5 points each: Hemogram (to assess anemia, leukocytosis, thrombocytopenia, or polycythemia) → **0.5 points** Biochemistry panel (to evaluate renal function, liver function, and metabolic status) → **0.5 points** Coagulation tests (PT, aPTT, INR) (to assess clotting status and risk of coagulopathy) → **0.5 points** D-dimer (for ruling out low-risk PE) → **0.5 points** **Maximum possible score: 10 points**
Question 4: Which imaging studies would you order to confirm your diagnosis? Explain your choices.	If the answer includes cardiac echocardiography (for assessing right ventricular strain, pulmonary hypertension, and indirect signs of PE) → **5 points** If the answer includes pulmonary computed tomography (CT) angiography (as the gold standard for confirming PE) → **3 points** If the answer includes pulmonary ventilation-perfusion scan (alternative in patients with contraindications to contrast-enhanced CT) → **1 point** If the answer includes chest X-ray (to rule out alternative diagnoses like pneumonia, pneumothorax, or pulmonary edema) → **1 point** **Maximum possible score: 10 points**
Case scenario continued Arterial blood gas analysis revealed respiratory alkalosis accompanied by hypoxemia. A bedside chest X-ray showed no signs of pulmonary edema or pneumonia, and neither Hampton's hump nor Westermark's sign was detected. The radiographic findings were evaluated as normal. Transthoracic echocardiography demonstrated right ventricular dysfunction and dilation, with a positive McConnell's sign. Additionally, tricuspid annular plane systolic excursion was found to be decreased, indicating right ventricular impairment. CT pulmonary angiography revealed massive bilateral thrombi in the right and left main pulmonary arteries, confirming the diagnosis of acute PE.	
Question 5: Which risk stratification tools would you use to assess this patient, and how do they contribute to PE management? Apply an appropriate tool to evaluate the severity of this patient's condition, calculate the corresponding score, and determine the risk category.	If the answer includes the pulmonary embolism severity index (PESI) as the risk stratification tool → **4 points** If the answer includes a correct PESI score calculation → **4 points** If the answer correctly classifies the patient into the appropriate risk category based on the PESI score → **2 points** (The patient's PESI score was calculated based on various clinical parameters. Because the patient was older than 65 years, 10 points were added to the score. Being female did not contribute any additional points. Systolic blood pressure below 100 mmHg contributed 30 points, while heart rate exceeding 110 bpm added 20 points. Additionally, a respiratory rate greater than 30 breaths per minute and oxygen saturation below 90% each contributed 20 points. The cumulative PESI score for this patient was calculated to be 140, categorizing her as high risk for PE). **Maximum possible score: 10 points**

Table 1. Continued

Questions	Predefined answer key
Question 6: Describe the stepwise clinical management of this patient in the emergency department. Outline the necessary interventions in order of priority, explaining the rationale for each step.	If the answer includes oxygen therapy for respiratory support → **2 points** If the answer includes IV fluid resuscitation (as appropriate for hemodynamic stabilization) → **2 points** If the answer includes vasopressors (for patients with shock or persistent hypotension) → **2 points** If the answer includes specific treatment targeting the diagnosis (anticoagulation therapy for PE) → **4 points** **Maximum possible score: 10 points**
Question 7: What is the appropriate dosing regimen and administration protocol for thrombolytic therapy in this patient?	If the answer includes alteplase (rtPA) at the standard dose of 100 mg over 2 hours → **6 points** If the answer includes the weight-based alternative alteplase regimen (0.6 mg/kg over 15 min, max 50 mg) → **4 points** If the answer includes the streptokinase regimen (250,000 IU loading dose over 30 min, followed by 100,000 IU/h for 12-24 hours) → **3 points** Alternative accelerated regimen: 1.5 million IU over 2 hours → **3 points** If the answer includes the urokinase regimen (4,400 IU/kg loading dose over 10 min, followed by 4,400 IU/kg/h for 12-24 hours) → **3 points** Alternative accelerated regimen: 3 million IU over 2 hours → **3 points** **Maximum possible score: 10 points**
Question 8: What clinical changes would you expect after thrombolytic therapy? How would you assess treatment success?	If the answer includes hemodynamic and clinical improvements (e.g., resolution of hypotension, improved oxygenation, symptom relief, reduced respiratory distress) → **4 points** If the answer includes echocardiographic changes (e.g., reduction in right ventricular strain, improved right ventricular function, decreased pulmonary artery pressure) → **4 points** If the answer includes ECG changes (e.g., resolution of S1Q3T3 pattern, T-wave normalization, reduction in right heart strain signs) or other supportive clinical markers → **2 points** **Maximum possible score: 10 points**
Case scenario continued Following thrombolytic therapy, the patient's hemodynamic status improved significantly. Her blood pressure stabilized at 110/70 mmHg, oxygen saturation increased to 95%, and heart rate decreased to 105 bpm. Given the initial high-risk presentation, the patient was admitted to the intensive care unit for close monitoring. Over the following days, her hemodynamic parameters remained stable, and repeat echocardiography showed improved right ventricular function. Once clinically stable, she was transferred to the inpatient ward for further management.	
Question 9: What tests should be performed before discharge, and what secondary prevention strategies and lifestyle modifications should be implemented?	If the answer includes lower extremity venous Doppler (to assess for residual DVT or ongoing thrombotic risk) → **2 points** If the answer includes a thrombophilia panel (to evaluate for inherited or acquired thrombophilia in selected patients) → **2 points** If the answer includes modifying or discontinuing oral contraceptive use (if applicable) → **2 points** If the answer includes prevention of prolonged immobilization (e.g., avoiding long sedentary periods, early ambulation, compression stockings if needed) → **2 points** If the answer includes regular exercise as part of secondary prevention → **1 point** If the answer includes weight loss (if the patient is overweight or obese) as a risk-reduction strategy → **1 point** **Maximum possible score: 10 points**
Question 10: How should this patient's discharge treatment plan be structured?	If the answer includes initiating low molecular weight heparin (e.g., dalteparin) for transition to oral anticoagulation (class 1, level A) → **3 points** If the answer includes long-term oral anticoagulation therapy (e.g., apixaban) (class 1, level A) → **3 points** If the answer includes the appropriate apixaban dosing regimen (first 7 days: 10 mg twice daily, then maintenance 5 mg twice daily) (class 1, level B) → **2 points** If the answer includes follow-up thoracic angiography to assess resolution of pulmonary artery thrombus (class 2a, level C) → **2 points** **Maximum possible score: 10 points**

This table evaluates the clinical reasoning of AI models across ten domains, including differential diagnosis, risk stratification, diagnostic testing, imaging, treatment, and follow-up. Responses adhering to guideline-based approaches are prioritized, emphasizing the importance of structured diagnosis, risk assessment, evidence-based treatment, and secondary prevention strategies

PT: prothrombin time, aPTT: activated partial thromboplastin time, INR: international normalized ratio, IV: intravenous, rtPA: recombinant tissue plasminogen activator, ECG: electrocardiography, DVT: deep vein thrombosis

Grok (X AI, USA): Included for its unique data sources, rapid growth via social media integration, and innovative approach under Elon Musk's X Corp.

Prompt Structure, Standardization and Data Collection

All models were accessed on March 10, 2025, using an anonymous user account to avoid any personalization or model adaptation based on prior interactions. Before initiating the formal assessment, a standardized role-conditioning prompt was applied to each model: "I am a physician. Please evaluate my questions as a physician and provide guideline-based clinical reasoning." This prompt was used to ensure a consistent baseline response style aligned with clinical decision-making.

Following this initial conditioning, the ten PE case–based questions were posed sequentially within the same conversation thread for each LLM, to maintain contextual continuity and to ensure that the models interpreted the scenario in a manner similar to real-world clinical reasoning.

To enhance reproducibility, the prompt engineering process was standardized across all LLMs. The same wording, order, and contextual flow were used for each model. No additional hints, sub-prompts, or clarifying questions were issued to the models beyond the predefined case scenario and the ten structured questions.

Each model received the same clinical vignette and follow-up questions without deviation. Importantly, no external information (e.g., computed tomography images) was uploaded or provided; instead, all radiological and laboratory findings were described textually to ensure consistent interpretation across platforms. All responses were saved immediately after generation and anonymized for evaluation.

Evaluation Process and Expert Review

The anonymized LLM responses and the predefined answer key were provided to two independent emergency medicine specialists, each with over 5 years of experience, who served as expert reviewers. Each expert was instructed to:

- Score each AI-generated response using the 10-point scoring system, based on adherence to evidence-based guidelines.

- In addition to quantitative scoring, reviewers were asked to comment qualitatively on clarity, structure, and clinical actionability. Although these qualitative dimensions were not included in the numerical scoring system, they were documented to enrich the interpretation of model performance.

Expert assessments were not limited to specific aspects of LLM-generated responses but rather aimed to capture broad observations and insights regarding their potential role in patient management.

Statistical Analysis

The collected expert ratings were compiled in Microsoft Excel and subsequently analyzed using IBM SPSS (latest version, IBM Corp., Armonk, NY, USA). Statistical analyses included descriptive (means, standard deviations, and frequency distributions) to summarize LLM performance. Normality was assessed using the Kolmogorov-Smirnov test. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) to ensure consistency in scoring across expert reviewers. A *P* value < 0.05 was considered statistically significant. All statistical procedures followed standard methodological guidelines to ensure the reliability and reproducibility of the findings.

RESULTS

For analytical purposes, the evaluation questions were classified into four categories: Diagnosis and Initial Evaluation (Q1–Q4), Risk Stratification and Prognosis (Q5), Management and Treatment (Q6–Q7), and Post-Treatment Assessment and Follow-Up (Q8–Q10).

ChatGPT-4o achieved the highest overall score (76), followed by Gemini (73.75), Grok (71.25), and DeepSeek-V2 (65). In the Diagnosis and Initial Evaluation category, all models exhibited comparable performance, with Q3 receiving the highest score (9.5 for both ChatGPT-4o and DeepSeek-V2). In the Risk Stratification and Prognosis category, DeepSeek-V2 obtained the lowest score of 5, whereas the other models obtained scores of 7. In the Management and Treatment category, Gemini outperformed ChatGPT-4o in Q7 (8 vs. 5); in Q6, ChatGPT-4o, Gemini, and Grok attained the highest possible score (10), whereas DeepSeek-V2 scored lower (7). In the Post-Treatment Assessment and Follow-Up category, ChatGPT-4o demonstrated superior performance, particularly in Q9, where it obtained the highest score (10), while Gemini had the lowest (6.5). The most pronounced performance discrepancy was observed in Q10, where ChatGPT-4o (4.5) and DeepSeek-V2 (4) outperformed Gemini (3) and Grok (1.5). Although ChatGPT-4o obtained the highest total score and DeepSeek-V2 obtained the lowest, no statistically significant differences were observed among the AI models in overall performance (H = 3.013, *P* = 0.390). The categorization of questions, the average scores assigned to each AI model, and the total scores are presented in Table 2.

Overall, the models' performance trends are variable: some models consistently achieve higher scores, while others demonstrate noticeable fluctuations. ChatGPT-4o reaches peak performance on certain questions (e.g., Q6 and Q9) but shows a significant decline in performance on Q7 and Q10. DeepSeek-V2 generally maintains a more stable but lower score range, with its lowest performance on Q5 followed by a gradual recovery in subsequent questions. Gemini demonstrates relatively stable performance, peaking at Q6 but exhibiting a decline towards the final questions. Grok follows a trend similar

Table 2. Performance comparison of medical AI models in acute pulmonary embolism assessment

	Question	ChatGPT-4o	DeepSeek-V2	Gemini	Grok
Diagnosis and initial evaluation	**Q1**	8	6.5	8	8
	Q2	6	6	6	6
	Q3	9.5	9.5	9.25	8.75
	Q4	8	8	8	8
Risk stratification and prognosis	**Q5**	7	5	7	7
Management and treatment	**Q6**	10	7	10	10
	Q7	5	6	8	6
Post-treatment assessment and follow-up	**Q8**	8	6	8	8
	Q9	10	7	6.5	8
	Q10	4.5	4	3	1.5
	Score	**76**	**65**	**73.75**	**71.25**

The table highlights the performance scores for each model across these categories, demonstrating variability in their responses and overall effectiveness in pulmonary embolism assessment

This table presents the evaluation of four AI models—ChatGPT-4o, DeepSeek-V2, Gemini, and Grok—across ten questions, covering domains such as diagnosis, risk stratification, management, treatment, and post-treatment follow-up

Q: question, AI: artificial intelligence to Gemini's but exhibits slightly greater variability, achieving higher scores on Q1 and Q6 while showing a sharp drop on Q10.

The interrater reliability between the two independent expert reviewers was assessed by calculating the ICC. The results indicated excellent agreement between the reviewers, with a single-measure ICC of 0.986 [95% confidence interval (CI): 0.975–0.992; $P < 0.001$] and an average-measure ICC of 0.993 (95% CI: 0.987–0.996; $P < 0.001$). The distribution of the average scores assigned by the expert reviewers to the AI-generated responses is visually represented in Figure 1.

Table 3 provides a qualitative summary of the expert reviewers' feedback, highlighting the strengths and weaknesses of each AI model. These findings underscore the variability in AI models' performance in clinical decision-making, demonstrating that while some models excel in structured and evidence-based responses, others offer advantages in risk stratification, accessibility, or practical applicability.

DISCUSSION

This study aimed to evaluate the performance of four popular AI LLMs in clinical decision-making based on the 2019 ESC guidelines for the management of PE. Investigating the potential of AI systems as clinical decision-support tools is important for accelerating and enhancing healthcare professionals' decision-making. The results demonstrated that while each AI model exhibited strong performance in specific domains, no model showed a clear overall superiority. ChatGPT-4o achieved the highest total score, whereas DeepSeek-V2 received the lowest total score. However, the varying performances of different AI models across clinical decision-making scenarios provide valuable insights into the strengths and limitations of each system.

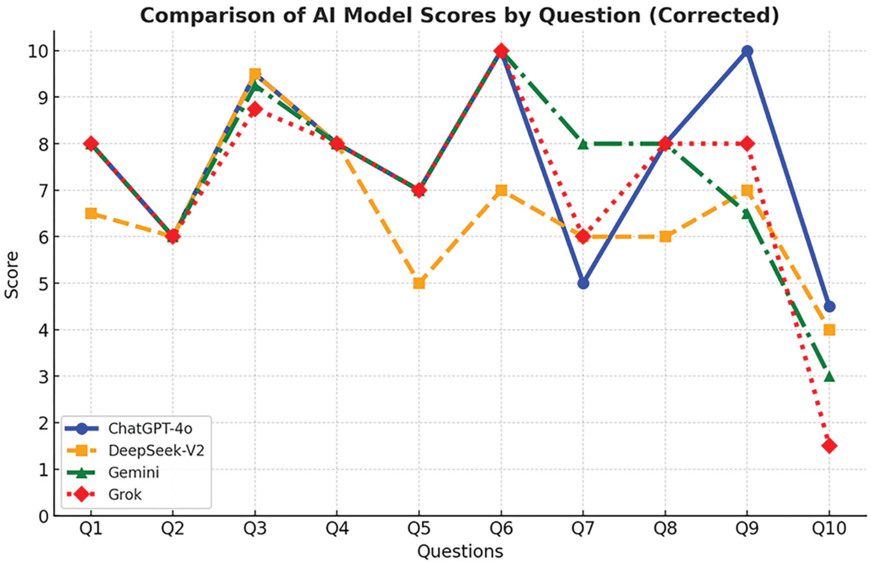

Figure 1. Comparison of AI models' scores on different questions

Comparison of AI models' scores on individual questions related to acute pulmonary embolism management. The line graph illustrates the performance of four AI models—ChatGPT-4o, DeepSeek-V2, Gemini, and Grok—across 10 structured questions (Q1–Q10). Each model's response was scored based on clinical accuracy, guideline adherence, clarity, and completeness

Q: question, AI: artificial intelligence

Table 3. Expert review of AI models in acute pulmonary embolism management

AI models	The expert reviewers' feedback
ChatGPT	ChatGPT was recognized for its comprehensive responses that adhered to guidelines, effectively addressing PE management in a structured, systematic manner. However, its emphasis on individualized treatment approaches was noted to be insufficient
DeepSeek	DeepSeek demonstrated notable proficiency in risk stratification and in clinical decision support, particularly in diagnostic test selection and stepwise management. Nevertheless, its explanations of thrombolytic therapy and post-discharge follow-up were deemed insufficient and required further elaboration
Gemini	Google Gemini provided fluent, accessible explanations consistent with general clinical practice, yet it lacked sufficient depth in risk stratification and in certain laboratory investigations
Grok	Grok was recognized for its practical and point-of-care management recommendations; however, its lack of a systematic approach and failure to provide direct references to established guidelines were identified as major limitations

Expert review of various AI models used in the management of acute pulmonary embolism. The table summarizes expert reviewers' evaluations regarding each model's strengths and limitations in areas such as diagnostic support, adherence to guidelines, risk stratification, and individualized treatment recommendations
AI: artificial intelligence, PE: pulmonary embolism

The variability in model performance likely reflects a combination of factors, including differences in model design, training objectives, and the general composition of training data; however, the exact sources and weighting of medical versus non-medical content are not publicly disclosed for these systems. Therefore, any explanation at the level of specific data sources or internal optimization strategies remains speculative. In this context, our findings are best interpreted as empirical evidence that each LLM exhibits domain-specific strengths and weaknesses, rather than as direct reflections of their proprietary training pipelines.

The literature encompasses a wide range of studies on the use of AI-based systems in the healthcare sector. It emphasizes that the use of AI in the medical field is rapidly increasing and may significantly enhance the efficiency of clinical decision support systems.[3,7-9] Our findings are consistent with previous studies showing that LLMs can generate clinically relevant, structured answers. However, the domain-specific strengths and weaknesses observed in each model underscore that LLMs should not be assumed to perform uniformly across all phases of patient management. From a clinical perspective, our results suggest that different LLMs may be better suited to particular components of PE management. For instance, DeepSeek-V2 demonstrated relative strength in diagnostics, Gemini performed strongly in treatment planning, and ChatGPT-4o excelled in post-treatment follow-up and in interpreting structured guidelines. These findings indicate that physicians should be aware of model-specific performance patterns rather than relying on a single system for all components of clinical care.

AI technology is evolving rapidly and becoming increasingly integrated into healthcare.[8] Continuous updates enhance AI performance in medical evaluation, as demonstrated in a gastroenterology study in which AI effectively managed real-world, guideline-based patient scenarios.[10] Studies investigating the medical reliability of publicly available LLMs suggest that these systems generate clinically relevant and guideline-compliant outputs. Our results support this conclusion by demonstrating that LLMs may provide reliable assistance in certain tasks, yet they also reveal that LLM-generated responses remain vulnerable to omissions, insufficient detail, or incomplete integration of risk stratification criteria in other tasks. Importantly, none of the evaluated models provided fully comprehensive or consistently guideline-aligned recommendations, reinforcing that LLMs should function as supportive tools rather than autonomous decision-makers. In a study evaluating the accuracy and clarity of patient education provided by three AI models on atrial fibrillation and cardiac implantable electronic devices, the responses generated by the AI systems were, on average, over 90% accurate and understandable.[3] In another AI study focusing on patient education regarding hypertension, the responses provided by ChatGPT to 100 questions were found to be appropriate and accurate in 92.5% of the cases.[7] In another study comparing two versions of ChatGPT in answering questions related to heart failure, GPT-3.5 achieved over 94% accuracy, while GPT-4 scored 100% across all 107 questions, demonstrating the high accuracy of these AI models in clinical decision-making for heart failure.[9] Our study supports these findings and provides valuable insights into how AI models can be applied to the management of critical medical conditions, such as PE.

Applying AI to PE assessment within a guideline-based framework offers valuable insight into its potential role in patient management. Additionally, studies examining healthcare professionals' perspectives on AI adoption indicate that these models could function as reliable clinical decision-support tools.[11,12] Beyond publicly available AI models, machine learning algorithms have been developed to improve acute PE diagnosis through electrocardiogram (ECG)-based analysis, further expanding AI's role in clinical medicine. In a study evaluating 1,014 ECGs obtained from patients admitted to the emergency department who underwent pulmonary computed tomography angiography for suspected PE, the AI model demonstrated greater specificity for detecting PE than commonly used prediction rules. The AI model achieved 100% specificity and 50% sensitivity.[13] Our study is consistent with these findings and provides important data on how AI models can be used to manage critical medical conditions such as PE. In this context, ChatGPT-4o was observed to produce highly accurate results, demonstrating a structured approach aligned with medical guidelines. However, some AI models, especially DeepSeek-V2, were found to produce less accurate responses or to provide insufficient coverage of specific treatment steps.

Given the life-threatening nature of PE, AI appears promising for real-world applications in emergency medicine. Integrating AI into medical decision-support systems may help mitigate healthcare workforce shortages, particularly in resource-limited settings.[14] While this study presents findings similar to those of those reported in the literature, it also highlights certain differences. For example, earlier studies have emphasized that AI generally contributes to diagnostic and therapeutic processes, though in some cases, clinical decision-making still requires significant human oversight.[8-10] In our study, the shortcomings of DeepSeek-V2, particularly in risk stratification and treatment processes, reveal notable limitations in the extent to which AI can be integrated into medical decision-making.

In this study, each AI model showed distinct strengths and weaknesses across different aspects of PE management. Below is a summary of their performance based on expert evaluations and guideline adherence:

ChatGPT-4o achieved the highest overall score, excelling in diagnosis and initial assessment. However, its limited ability to provide personalized treatment plans highlights the need for improved patient-specific adaptability—especially important in emergency settings.

DeepSeek-V2 showed stable but generally low performance, particularly in risk stratification and lab interpretation. While it performed well on some individual questions, its limited coverage of treatment and follow-up reduces its clinical utility.

Gemini stood out in treatment management, offering well-structured therapeutic recommendations. However, it had difficulty with risk assessment and interpretation of lab results. These gaps may be addressed in future updates, given Google's resources and AI capabilities.

Grok provided practical, point-of-care suggestions but lacked a clear structure and direct references to guidelines, thereby limiting its scientific reliability. Its wide reach via social media is a strength, but better integration with evidence-based content is needed to enhance clinical applicability.

In our study, AI models demonstrated varying strengths and limitations in the management of PE, as reflected in expert evaluations. ChatGPT was praised for providing comprehensive management due to its structured and guideline-compliant approach; however, its lack of focus on individualized treatment was noted as a drawback. DeepSeek performed well in risk stratification and diagnostic test selection, enhancing clinical decision-making, but its limited coverage of thrombolytic therapy and post-discharge follow-up represented a limitation. Gemini provided clear, accessible explanations and aligned with general clinical practice; nevertheless, it lacked depth in risk assessment and in certain laboratory evaluations critical to PE management. Grok was found to be useful in offering practical, point-of-care recommendations; however, the absence of a structured approach and of direct references to guidelines were cited as major limitations. These findings highlight that although AI models can provide valuable clinical insights, their performance varies across different aspects of PE management, underscoring the need for model refinement, guideline integration, and a more personalized approach.

For future research, more robust evaluation strategies are recommended, including multi-scenario studies, comparative assessments across different sets of clinical guidelines, and investigations of combined or ensemble approaches that leverage the strengths of multiple LLMs. Additionally, prospective studies examining how LLM support influences physician decision-making accuracy, workflow efficiency, and patient outcomes would provide meaningful insights into real-world applicability. Frameworks conceptualized as "physician–LLM collaboration models" may also help define safer and more effective integration pathways.

Study Limitations

Several considerations should guide the interpretation of this study. The analysis was based on a single clinical scenario, which enabled standardized comparison but naturally limits the breadth of clinical contexts to which the findings can be generalized. Different presentations may challenge LLMs in distinct ways.

Although all models were evaluated on the same day and under uniform conditions, LLMs evolve rapidly, and their performance reflects a moment in time rather than a stable characteristic. Future updates may alter their reasoning patterns.

As with all generative models, the risk of factual distortion or overconfident statements remains inherent. Our structured scoring system reduced this risk but could not completely eliminate it.

Finally, while quantitative scoring allowed for reproducible evaluation, the absence of dedicated qualitative metrics—such as omission tracking or actionability—represents a conceptual limitation that future studies may address.

CONCLUSION

In our study, each model demonstrated distinct strengths across the diagnostic, treatment, and follow-up processes; however, performance fluctuations were particularly notable in areas such as personalized patient management and risk assessment. Although AI models show promising potential as clinical decision support tools, they should be further trained with real-world patient data to enhance adherence to clinical guidelines and better align with the principles of personalized medicine. In this context, they should be developed not to replace physicians but to support clinical decision-making.

Ethics

Ethics Committee Approval: Since our research did not involve any human participants and did not include any patient records, no application to an ethics committee was submitted. Nevertheless, throughout all stages of the study, the principles of scientific research and publication ethics were strictly observed. Should any additional information or clarification be required, we are prepared to provide detailed information.

Informed Consent: This study is a simulation-based investigation and does not involve human participants or the use of real patient data.

Artificial Intelligence Usage Statement

Artificial intelligence (AI) models, including ChatGPT-4o, Gemini, Grok, and DeepSeek-V2, were used solely as subjects in the comparative evaluation. No generative AI tool was used to create or edit the manuscript content. All analyses, interpretations, and manuscript writing were performed by the authors.

Footnotes

Authorship Contributions

Surgical and Medical Practices: Ö.F.K., H.E.K., M.E.Ö., Y.G., Concept: Ö.F.K., H.E.K., M.E.Ö., B.Y., Design: Ö.F.K., H.E.K., M.E.Ö., B.Y., Data Collection or Processing: Ö.F.K., H.E.K., Ö.H.S., M.E.Ö., Analysis or Interpretation: Ö.F.K., H.E.K., Ö.H.S., M.E.Ö., Y.G., Literature Search: Ö.F.K., H.E.K., Ö.H.S., Y.G., Writing: Ö.F.K., H.E.K., B.Y.

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. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered Chatbots in medical education: potential applications and implications. *Cureus*. 2023;15(8):e43271. **[Crossref]**
2. Niko MM, Karbasi Z, Kazemi M, Zahmatkeshan M. Comparing ChatGPT and Bing, in response to the Home Blood Pressure

Monitoring (HBPM) knowledge checklist. *Hypertens Res*. 2024;47(5):1401-1409. **[Crossref]**

3. Hillmann HAK, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace*. 2023;26(1):euad369. **[Crossref]**
4. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, et al. The future of patient education: AI-driven guide for type 2 diabetes. *Cureus*. 2023;15(11):e48919. **[Crossref]**
5. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *J Med Syst*. 2024;48(1):22. **[Crossref]**
6. Konstantinides SV, Meyer G, Becattini C, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). *Eur Heart J*. 2020;41(4):543-603. **[Crossref]**
7. Almagazzachi A, Mustafa A, Eighaei Sedeh A, et al. Generative artificial intelligence in patient education: ChatGPT takes on hypertension questions. *Cureus*. 2024;16(2):e53441. **[Crossref]**
8. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: trends in artificial intelligence for biotechnology. *New Biotechnol*. 2023;74:16-24. **[Crossref]**
9. King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. *Heart Lung Circ*. 2024;33(9):1314-1318. **[Crossref]**
10. Sciberras M, Farrugia Y, Gordon H, et al. Accuracy of information given by ChatGPT for patients with inflammatory bowel disease in relation to ECCO guidelines. *J Crohns Colitis*. 2024;18(8):1215-1221. **[Crossref]**
11. Vearrier L, Derse AR, Basford JB, Larkin GL, Moskop JC. Artificial intelligence in emergency medicine: benefits, risks, and recommendations. *J Emerg Med*. 2022;62(4):492-499. **[Crossref]**
12. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak*. 2023;23(1):73. **[Crossref]**
13. Valente Silva B, Marques J, Nobre Menezes M, Oliveira AL, Pinto FJ. Artificial intelligence-based diagnosis of acute pulmonary embolism: development of a machine learning model using 12-lead electrocardiogram. *Rev Port Cardiol*. 2023;42(7):643-651. English, Portuguese. **[Crossref]**
14. Long P, Lu L, Chen Q, Chen Y, Li C, Luo X. Intelligent selection of healthcare supply chain mode - an applied research based on artificial intelligence. *Front Public Health*. 2023;11:1310016. **[Crossref]**