Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(1):116-21

Evaluating a large language model's accuracy in CTPA image interpretation for acute pulmonary embolism

Banu Arslan

Başakşehir Çam and Sakura City Hospital, Department of Emergency Medicine, İstanbul, Türkiye

Received 21 June 2025; Accepted 30 August 2025
Available online 20 February 2026 with doi: 10.5455/medscience.2025.06.166

Abstract

Pulmonary embolism (PE) is a major cause of cardiovascular mortality. Computed tomography pulmonary angiography (CTPA) is the gold standard for definitive diagnosis; however, image interpretation can be delayed in busy or under-resourced emergency departments. Large language models (LLMs) such as ChatGPT-4 Turbo, which now accept images, may offer scalable decision support, but their diagnostic performance on CTPA has not yet been defined. We retrospectively assembled 200 de-identified single-slice CTPA images from PubMed Central (162 positive, 38 negative for PE). After anonymization, each image was uploaded individually to ChatGPT-4 Turbo. Ground-truth labels were obtained from source case reports. Sensitivity, specificity, accuracy, predictive values, and Cohen's kappa for anatomic sub-typing were calculated. ChatGPT identified 150 true positives, 12 false negatives, 28 false positives, and 10 true negatives. Sensitivity was 92.6%, specificity 26.3%, and overall accuracy 80%. Positive and negative predictive values were 84.3% and 45.5%, respectively. Agreement on embolus location was fair (κ=0.25) with 51% correct sub-typing. ChatGPT-4 Turbo detected most emboli but generated many false alarms and misclassified half of the anatomic locations. These data position the model as a useful triage aid to flag CTPAs for expedited human review, rather than a stand-alone interpreter.

Keywords: ChatGPT, artificial intelligence, computed tomography pulmonary angiography, pulmonary embolism, emergency

Introduction

Pulmonary embolism (PE) is a potentially life-threatening condition resulting from the obstruction of pulmonary arteries by thrombotic material, most commonly originating from deep vein thrombosis. It represents a major cause of cardiovascular morbidity and mortality worldwide [1]. Clinical presentations range from subtle symptoms to sudden cardiac arrest. Rapid and accurate diagnosis is critical, as delays in initiating treatment can lead to significant adverse outcomes. PE was diagnosed using imaging modalities such as computed tomography pulmonary angiography (CTPA) and ventilation–perfusion scintigraphy [2]. CTPA is the current gold standard imaging modality for detecting PE due to its high sensitivity and specificity in visualizing intravascular filling defects [3].

Despite the diagnostic capabilities of CTPA, interpretation remains time-sensitive and resource-dependent, particularly in busy emergency departments or facilities lacking specialized radiologists. Artificial intelligence (AI), especially large language model (LLM)-based tools like ChatGPT, has recently emerged as a promising adjunct in medical image interpretation. These models can process visual data, extract clinically relevant features, and generate structured responses [4], potentially improving diagnostic workflows and reducing interobserver variability [5,6]. In this study, we aimed to evaluate the accuracy of ChatGPT, a widely used generative AI model, in identifying the presence and anatomical location of PE on CTPA images. By assessing both detection performance and agreement with ground truth classification, we sought to explore the feasibility of incorporating ChatGPT into radiological evaluation processes in emergency departments.

Previous AI applications in radiology have primarily focused on deep learning models trained specifically on medical datasets. While these models may have demonstrated promising results, their deployment often requires specialized infrastructure, technical expertise, and continual retraining. In contrast, ChatGPT offers an accessible, general-purpose AI interface

CITATION
Arslan B. Evaluating a large language model's accuracy in CTPA image interpretation for acute pulmonary embolism. Med Science. 2026;15(1):116-21.

Corresponding Author: Banu Arslan, Başakşehir Çam and Sakura City Hospital, Department of Emergency Medicine, İstanbul, Türkiye
Email: dr.banuarslan@gmail.com

DOI: 10.5455/medscience.2025.06.166

capable of handling visual inputs without domain-specific pretraining [7]. This could provide a practical and scalable tool for supporting clinicians in diverse settings, including under-resourced healthcare environments. However, its performance in diagnostic image interpretation remains underexplored, particularly for high-stakes conditions such as PE.

In this study, we assessed ChatGPT's ability to detect PE and correctly identify its anatomical classification (saddle, central, segmental, subsegmental). By comparing its outputs to reference annotations and calculating key diagnostic metrics—including sensitivity, specificity, accuracy, and interrater agreement—we aimed to provide early insights into the potential role of LLM-based generative AI in augmenting the radiological assessment of PE.

Material and Methods

Study Design

This study was as a retrospective, cross-sectional analysis to evaluate the accuracy of ChatGPT in detecting PE on CTPA images. The study was conducted in accordance with the STARD guidelines for reporting diagnostic accuracy. Institutional Review Board (IRB) approval was not required, as the study utilized publicly available and de-identified CTPA images. The overall study workflow, including image selection, anonymization, AI interpretation, and comparison with the gold standard, is shown in Figure 1.

Dataset

Study data were obtained from PubMed Central (PMC), a free full-text archive maintained by the U.S. National Institutes of Health (NIH). At the time of access, 14.823 figures were labeled as depicting PE. After careful assessment, only CTPA images were selected based on the following criteria: high image quality, axial slice orientation, mediastinal window setting, and absence of excessive annotations or visual markers such as arrows or text overlays. Only images from individual patients were included, and those with low resolution, non-axial views, or obstructive labels were excluded. Following this screening process, 200 CTPA images from 200 unique patients were included in the study.

ChatGPT

ChatGPT (OpenAI, San Francisco, CA, USA), a natural language processing (NLP) model, was employed in this study through ChatGPT Plus with GPT-4 Turbo. GPT-4-turbo is the AI model behind ChatGPT as of 2025. Compared to the free version, this model offers faster performance, greater efficiency, and enhanced capabilities—including handling images, files, and code. Additionally, it can access to the continuously updated tools such as Python, image generation, and web browsing, depending on the user's settings [8].

Prior to uploading the images for interpretation, we ensured that all CTPA images were fully anonymized, with no identifiable information or visual indicators such as asterisks, arrows, or color overlays. A new project was initiated, and ChatGPT provided with the study context. The model was then prompted to answer the following two questions for each image:

1. "Is there a pulmonary embolism present?"
2. "If yes, what is the likely location (e.g., saddle, central, segmental, subsegmental)?"

The model allows uploading up to10 images at a time. However we uploaded each image individually, and the response was recorded before proceeding to the next. Figure 2 illustrates the interaction between ChatGPT and the user, showing a saddle pulmonary embolism on a single-slice CTPA along with the corresponding interpretation generated by ChatGPT-4 Turbo [9]. This process was repeated until all 200 images were interpreted. The model's answers were systematically recorded and included in the study dataset. Only the initial response provided by ChatGPT was considered, and the same questions were not repeated to avoid response variation.

Statistical Analysis

Descriptive statistics were used to analyze the collected data and assess the effectiveness of ChatGPT in image interpretation. Categorical variables were compared using the chi-square test. All statistical analyses were performed using Microsoft Excel (version 14.7.1; Microsoft Corp, Redmond, WA) and IBM SPSS Statistics (version 28; IBM Corp, Armonk, NY, USA). A P value of <0.05 was considered statistically significant.

Result

This study included 200 CTPA images evaluated for PE. Of these, 162 (81%) images demonstrated the presence of PE, while 38 (19%) showed no evidence of occlusion in the pulmonary arteries. Based on anatomical distribution, 27 cases (17%) were classified as saddle PE, 87 (54%) as central, 36 (22%) as segmental, and 12 (7%) as subsegmental embolism. Regarding laterality, 30% of cases were confined to the right lung, 19% to the left lung, and 52% involved both lungs.

The model identified 150 true positives, 10 true negatives, 28 false positives, and 12 false negatives. Accordingly, its sensitivity was 92.6% (95% CI: 88.6–96.6), specificity was 26.3% (95% CI: 12.3–40.3), and overall accuracy was 80.0% (95% CI: 74.5–85.5). The positive predictive value (PPV) was calculated as 84.3% (95% CI: 78.9–89.6), while the negative predictive value (NPV) was 45.5% (95% CI: 24.6–66.3) (Table 1). The overall accuracy of ChatGPT in identifying the specific location of PE was approximately 51.2% (95% CI: 43.5%–58.9%).

The interrater reliability (IRR) between the ground truth and ChatGPT for the embolism classification was also calculated. The Cohen's kappa coefficient was 0.25, indicating fair agreement (Table 2).

DOI: 10.5455/medscience.2025.06.166

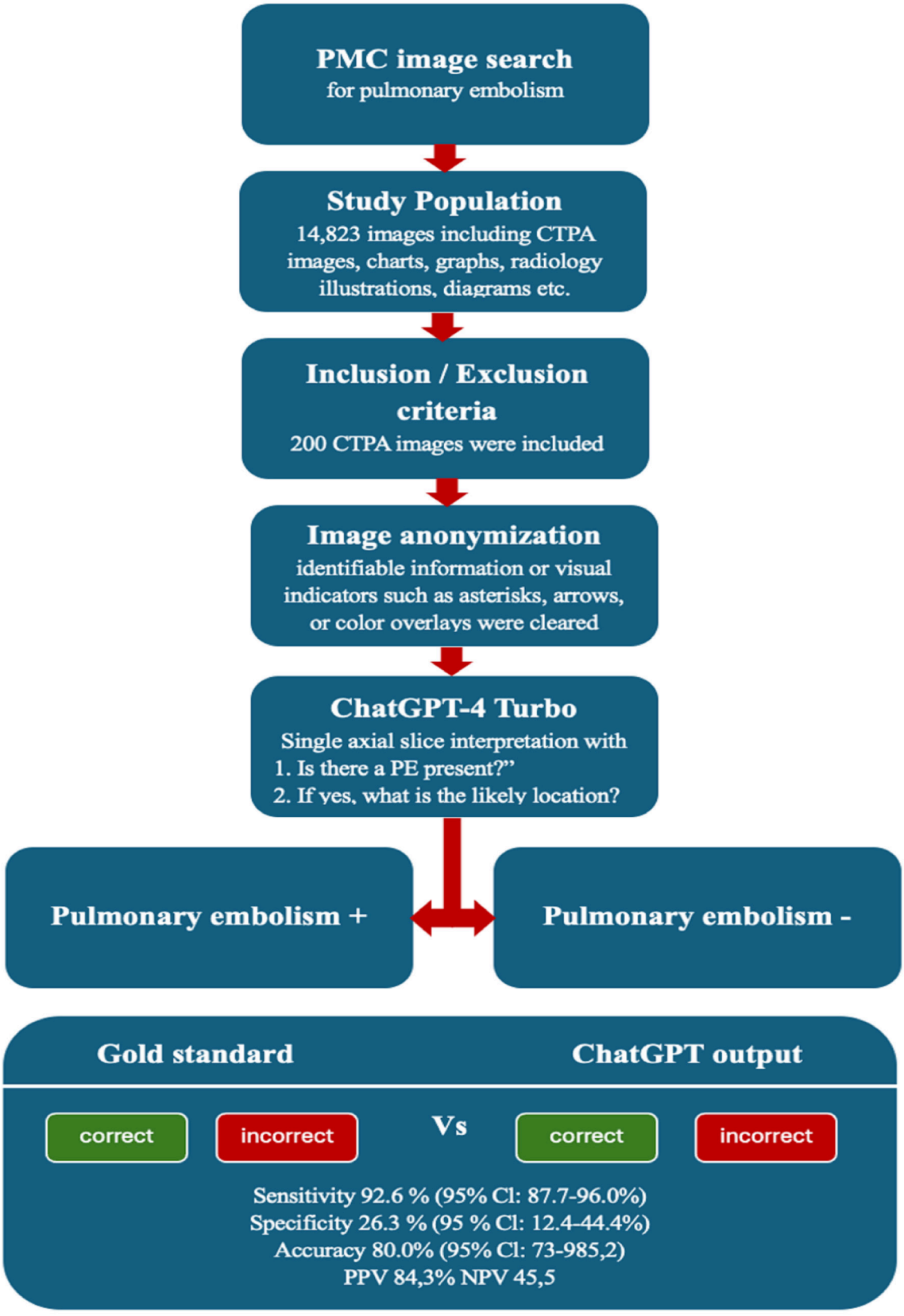
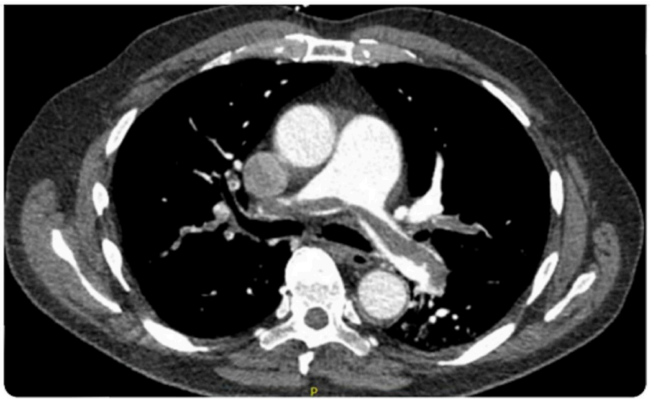


Figure 1. Flowchart illustrating the study design for evaluating ChatGPT-4 Turbo's performance in detecting pulmonary embolism on CTPA images, including image selection, anonymization, interpretation, and comparison with the gold standard

DOI: 10.5455/medscience.2025.06.166



CTPA Interpretation:

1. Pulmonary Embolism Present?
   ✅ Yes
2. Likely Location of Embolism:
   Saddle – There is a large filling defect clearly straddling the bifurcation of the main pulmonary artery and extending into both the right and left pulmonary arteries, which is characteristic of a saddle embolus.

You can send the next image whenever you're ready!

Figure 2. Example of a saddle pulmonary embolism on single-slice CTPA and the corresponding interpretation generated by ChatGPT-4 Turbo. The model identified the filling defect at the bifurcation of the main pulmonary artery and classified it as a saddle pulmonary embolism, consistent with the reference standard. The AI-generated text output included a brief interpretation ("There is a filling defect at the bifurcation of the main pulmonary artery consistent with a saddle pulmonary embolism"), which aligned with the ground truth from the source case report; CTPA image is retrieved from PubMed Central [9]

Table 1. Diagnostic performance metrics of ChatGPT in detecting pulmonary embolism on CTPA images

Metric	Sensitivity	Specificity	Accuracy	PPV*	NPV*
Value (%)	92.6	26.3	80.0	84.3	45.5
95% CI	88.6–96.6	12.3–40.3	74.5-85.5	78.9–89.6	24.6–66.3

PPV: positive predictive value, NPV: negative predictive value

Table 2. Confusion matrix showing agreement between ground truth and ChatGPT in anatomic classification of pulmonary embolism on CTPA images (n=150 positive cases)

Interpreters		ChatGPT				TOTAL
		Saddle	Central	Segmental	Subsegmental	
Ground Truth	Saddle	11	14	1	0	26
	Central	25	53	7	0	85
	Segmental	5	9	18	0	32
	Subsegmental	2	3	1	1	7
TOTAL		43	79	27	1	150

Cohen's kappa: 0.25, indicating fair agreement)

DOI: 10.5455/medscience.2025.06.166

Discussion

Recent studies have evaluated the accuracy of ChatGPT in interpreting various types of medical imaging such as chest X-ray, head CT, diffusion MRI and abdominal MRI [10-13]. However, data regarding its accuracy is still limited. Lacaita et al. evaluated 257 chest radiographs and reported a moderate overall accuracy of 66.2% (95% CI, 60.2–71.7) [10]. Ostrovsky et al. found that the model reliably recognized normal chest X-rays, but its diagnostic strength varied by pathology: for pneumonia, sensitivity and specificity were 76.2% and 93.7%, respectively, whereas for atelectasis they declined to 44.4% and 89.1% [14]. In chest CT interpretation, Dehdab et al. observed an overall accuracy of 56.8% with ChatGPT-4V [15].

To our knowledge, this is the first study to evaluate ChatGPT's accuracy in detecting PE on CTPA images. Our findings demonstrate that ChatGPT-4 Turbo achieved a high sensitivity (92.6%) but a markedly low specificity (26.3%) for the detection of PE on CTPA images. The low specificity observed in our study likely reflects several methodological and model-related factors. First, only single axial CTPA slices were analyzed, which deprived the model of volumetric context that radiologists routinely use to differentiate true intraluminal thrombus from normal anatomical structures or artifacts. Without multi-slice correlation, common mimics such as beam-hardening, motion artifacts, or vascular branching patterns may be misclassified as emboli [16]. Second, ChatGPT-4 Turbo is a general-purpose model without domain-specific fine-tuning on curated CTPA datasets. This lack of targeted radiology training likely limits its ability to confidently exclude pathology, predisposing it to "overcall" suspicious findings. Third, because the model was evaluated without integration of clinical context or laboratory results, it lacked important discriminative cues that radiologists often use to rule out non-pathologic findings.

This low specificity may increase the number of false-positive detections. In a clinical setting, this could lead to unnecessary escalations, increased interpretation workload, and potential "alert fatigue" among clinicians, particularly in busy or resource-limited emergency departments. Alert fatigue, a well-documented phenomenon in clinical decision support systems, can reduce clinician responsiveness over time and undermine the intended safety benefits [17]. For a triage application, the trade-off between sensitivity and specificity must therefore be carefully managed. On the other hand, these findings represent an early step in applying this cutting-edge technology to complex radiological tasks. As multimodal AI models evolve, particularly with advancements in video and volumetric image interpretation [18], ChatGPT's performance in detecting and localizing PE on CTPA is expected to improve. From this perspective, our results suggest that, without specialized training, ChatGPT functions more as an over-inclusive screening tool than a definitive interpreter.

ChatGPT achieved only fair agreement with the reference standard for anatomic sub-typing (κ=0.25; 51 % accuracy). While anatomical burden and distribution of PE may contribute to in risk stratification and treatment selection [19], acute management still hinges on the patient's early mortality risk profile; namely hemodynamic stability, right-ventricular dysfunction, troponin elevation, and clinical scores such as PESI [20,21]. Therefore, while the model reliably detects most emboli, its limited precision in localizing means it cannot replace expert interpretation, and radiologist confirmation remains essential.

The ethical and regulatory implications of deploying general-purpose AI in clinical care warrant careful consideration. Unlike U.S. Food and Drug Administration or European Conformity cleared medical imaging software, ChatGPT-4 Turbo is not specifically approved for diagnostic purposes. Although not originally intended for medical use, it has been widely adopted due to its intuitive, barrier-free user interface, capacity to generate human-like text, and broad domain knowledge across diverse subjects [22]. Many users are willing to consult ChatGPT for health-related questions; however, its use raises ethical concerns, including privacy, accuracy, and bias [23]. Furthermore, medico-legal responsibility in the event of diagnostic error remains unclear, with potential liability resting with the developer, the healthcare institution, or the interpreting clinician, depending on jurisdiction.

This study has several important limitations. First, single axial images were analyzed rather than full volumetric series, depriving the model of contextual information radiologists routinely use. Second, ground-truth labels derived from published case reports may carry residual bias. Third, the dataset size (n=200) limits statistical precision, particularly for subsegmental emboli. Another limitation is the relatively small number of PE-negative cases (n=38), which reduces the precision of specificity and NPV estimates. With such a small denominator, each false positive alters the specificity by approximately 2.6 percentage points, contributing to a wide confidence interval (95% CI: 12.3–40.3) and reduced precision of this metric. Future work should therefore explore (i) multi-slice or 3-D inputs, (ii) fine-tuning with curated CTPA datasets, and (iii) prospective trials measuring impact on time-to-diagnosis and clinical outcomes. Until such advances are realized—and regulatory pathways clarified—ChatGPT is best viewed as an adjunctive, not autonomous, tool for PE assessment, capable of accelerating workflow but requiring vigilant human oversight.

Conclusion

Our pilot evaluation shows that ChatGPT-4 Turbo can detect pulmonary embolism on single-slice CTPA with high sensitivity but low specificity. Although the model correctly flagged most emboli, its location agreement with expert annotation indicated limited reliability for precise anatomic sub-typing. Taken together, these findings position ChatGPT as a potential triage aid—capable of prioritizing studies that warrant urgent human review—rather than a stand-alone diagnostic reader.

DOI: 10.5455/medscience.2025.06.166

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
Ethical committee approval was unnecessary for this study.

References

1. Vyas V, Sankari A, Goyal A. Acute pulmonary embolism. https://www.ncbi.nlm.nih.gov/books/NBK560551/ access date 11.12.2024
2. Gegin S, Karakaya T, Arslan Aksu E, et al. The distribution of hereditary risk factors in patients with pulmonary thromboembolism without identifiable acquired risk factors. Duzce Medical Journal. 2025;27:69-74.
3. Bukhari SMA, Hunter JG, Bera K, et al. Clinical and imaging aspects of pulmonary embolism: a primer for radiologists. Clin Imaging. 2025;117:110328.
4. Partha Pratim Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;3:121-154.
5. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: The advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus. 2023;15:e41435.
6. Kaboudi N, Firouzbakht S, Shahir Eftekhar M, et al. Diagnostic accuracy of ChatGPT for patients' triage; A systematic review and meta-analysis. Arch Acad Emerg Med. 2024;12:e60.
7. GPT-4, https://openai.com/research/gpt-4 2023
8. New models and developer products announced at DevDay, https://openai.com/index/new-models-and-developer-products-announced-at-devday/?utm_source=chatgpt.com access date 06.18.2025
9. Kristensen AMD, Rosberg V, Juel J, Pareek M. Conservatively managed saddle pulmonary embolism. Clin Case Rep. 2019;7:1259-60.
10. Lacaita PG, Galijasevic M, Swoboda M, et al. The accuracy of ChatGPT-4o in interpreting chest and abdominal X-Ray images. J Pers Med. 2025;15:194.
11. Abrigo JM, Ko KL, Chen Q, et al. Artificial intelligence for detection of intracranial haemorrhage on head computed tomography scans: diagnostic accuracy in Hong Kong. Hong Kong Med J. 2023;29:112-20.
12. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. Diagn Interv Radiol. 2025;31:187-95.
13. Elek A, Ekizalioğlu DD, Güler E. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. Diagn Interv Radiol. 2025;31:196-205.
14. Ostrovsky AM. Evaluating a large language model's accuracy in chest X-ray interpretation for acute thoracic conditions. Am J Emerg Med. 2025;93:99-102.
15. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. Jpn J Radiol. 2024;42:1168-77.
16. Hutchinson BD, Navin P, Marom EM, et al. Overdiagnosis of pulmonary embolism by pulmonary CT Angiography. AJR Am J Roentgenol. 2015;205:271-7.
17. Baseman JG, Revere D, Painter I, et al. Public health communications and alert fatigue. BMC Health Serv Res. 2013;13:295.
18. Maaz M, Rasheed H, Khan S, Khan F. Video-chatgpt: Towards detailed video understanding via large vision and language models. 2024;1:12585-602.
19. Vedovati MC, Germini F, Agnelli G, Becattini C. Prognostic role of embolic burden assessed at computed tomography angiography in patients with acute pulmonary embolism: systematic review and meta-analysis. J Thromb Haemost. 2013;11:2092-102.
20. Konstantinides SV, Meyer G. The 2019 ESC Guidelines on the Diagnosis and Management of Acute Pulmonary Embolism. Eur. Heart J. 2019;40:3453–5.
21. Leidi A, Bex S, Righini M, et al. Risk stratification in patients with acute pulmonary embolism: current evidence and perspectives. J Clin Med. 2022;11:2533.
22. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: A taxonomy and systematic review. Comput Methods Programs Biomed. 2024;245:108013.
23. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.