ORIGINAL ARTICLE

DOI: 10.4274/nkmj.galenos.2026.00921
Nam Kem Med J 2026;14(2):179-189

Evaluation of the Diagnostic Performance of Large Language Models in Distinguishing Pulmonary Embolism and Pulmonary Artery Sarcoma

Pulmoner Emboli ve Pulmoner Arter Sarkomunu Ayırt Etmede Büyük Dil Modellerinin Tanısal Performansının Değerlendirilmesi

Hadi SASANİ[1], Mehmet Ali ŞİMŞEK[2]

[1]*Tekirdağ Namık Kemal University Faculty of Medicine, Department of Radiology, Tekirdağ, Türkiye*
[2]*Bandırma Onyedi Eylül University Faculty of Engineering and Natural Sciences, Department of Software Engineering, Balıkesir, Türkiye*

ABSTRACT

Aim: Pulmonary embolism (PE) and pulmonary artery sarcoma (PAS) present with similar clinical symptoms but differ significantly in pathology and treatment. Accurate differentiation is critical yet challenging in clinical practice. This study aimed to evaluate the diagnostic efficacy of large language models (LLMs) in distinguishing PE from PAS, and to explore their potential as clinical decision support tools.

Materials and Methods: Eighteen cases with confirmed diagnoses of PE or PAS were assessed using three LLMs: DeepSeek V3, Gemini Flash 2.5, and ChatGPT-4o. Models were provided with basic clinical data, followed by advanced imaging and treatment information. The Role-Goal-Context framework was applied to standardize the input prompts.

Results: DeepSeek V3 achieved the highest accuracy in detecting PE during the preliminary diagnostic phase (88.9 %), while Gemini Flash 2.5 performed best in identifying PAS in the conclusive phase (22.2 %). ChatGPT-4o yielded more stable results under conditions of limited data availability.

Conclusion: LLMs show promise as supportive tools in differentiating PE from PAS when guided by structured prompts and expert oversight. Their utility is limited in rare or data-deficient scenarios. A hybrid model involving human expertise and LLM integration appears most effective in enhancing diagnostic precision and clinical decision-making.

Keywords: Pulmonary embolism, pulmonary artery sarcoma, large language models, diagnostic accuracy, clinical decision support

ÖZ

Amaç: Pulmoner emboli (PE) ve pulmoner arter sarkomu (PAS) benzer klinik semptomlarla ortaya çıkar, ancak patoloji ve tedavi açısından önemli farklılıklar gösterir. Klinik uygulamada doğru ayırıcı tanı koymak çok önemlidir, ancak zordur. Bu çalışma, PE ile PAS'ı ayırt etmede büyük dil modellerinin (LLM) tanısal etkinliğini değerlendirmek ve klinik karar destek araçları olarak potansiyellerini araştırmak amacıyla yapılmıştır.

Gereç ve Yöntem: PE veya PAS tanısı doğrulanmış 18 olgu, DeepSeek V3, Gemini Flash 2.5 ve ChatGPT-4o olmak üzere üç LLM kullanılarak değerlendirilmiştir. Modeller, temel klinik verilerle beslendikten sonra ileri görüntüleme ve tedavi bilgileriyle beslenmiştir. Giriş komutlarını standartlaştırmak için Rol-Hedef-Bağlam çerçevesi uygulanmıştır.

Bulgular: DeepSeek V3, ön tanı aşamasında PE'yi tespit etmede en yüksek doğruluğu elde ederken (%88,9), Gemini Flash 2.5 kesin aşamada PAS'ı tanımlamada en iyi performansı gösterdi (%22,2). ChatGPT-4o, sınırlı veri kullanılabilirliği koşullarında daha istikrarlı sonuçlar verdi.

Sonuç: LLM'ler, yapılandırılmış komutlar ve uzman gözetimi ile yönlendirildiğinde, PE'yi PAS'tan ayırt etmede destekleyici araçlar olarak umut vaat etmektedir. Nadir veya veri eksikliği olan senaryolarda kullanışlılıkları sınırlıdır. İnsan uzmanlığı ve LLM entegrasyonunu içeren hibrit bir model, tanı doğruluğunu ve klinik karar vermeyi geliştirmede en etkili görünmektedir.

Anahtar Kelimeler: Pulmoner emboli, pulmoner arter sarkomu, büyük dil modelleri, tanı doğruluğu, klinik karar desteği

Address for Correspondence: Asst. Prof. Mehmet Ali ŞİMŞEK, Bandırma Onyedi Eylül University Faculty of Engineering and Natural Sciences, Department of Software Engineering, Balıkesir, Türkiye
E-mail: msimsek@bandirma.edu.tr **ORCID ID:** orcid.org/0000-0002-6127-2195
Received: 17.12.2025 **Accepted:** 09.02.2026 **Publication Date:** 16.06.2026

Cite this article as: Sasani H, Şimşek MA. Evaluation of the diagnostic performance of large language models in distinguishing pulmonary embolism and pulmonary artery sarcoma. Nam Kem Med J. 2026;14(2):179-189

INTRODUCTION

Acute pulmonary embolism (PE) is a common and potentially lethal result of venous thromboembolic disease (VTE). D-dimer, a plasmin-derived degradation product, has a high sensitivity and negative predictive value in the diagnosis of VTE. As imaging modalities in the diagnosis of PE, lower extremity venous Doppler ultrasonography and echocardiography, as well as pulmonary computed tomography angiography (CTA) and scintigraphy, magnetic resonance angiography (MRA), and conventional pulmonary angiography. However, pulmonary CTA is the most crucial and widely utilized imaging modality for the diagnosis of suspected PE, with high sensitivity and specificity. A central filling defect in a vessel encircled by contrast material is one of the direct CT findings of acute PE[1].

However, pulmonary artery sarcoma (PAS) should also be considered in the differential diagnosis because it appears as a defect filling with a similar appearance to PE. PAS is a rare and malignant tumor that develops within the inner or middle layer of the pulmonary artery, with an estimated incidence rate ranging from 0.001-0.03%, with a male-to-female ratio of 1:2. It mainly occurs in middle-aged people, with an average onset age of approximately 50 years. The survival period of PAS is approximately 1.5 months for those who did not undergo surgery timely. If there is a suspicion of PAS in the presence of a lesion that does not improve or persists despite PE treatment, it should be evaluated with positron emission tomography/computed tomography (PET/CT). PAS typically presents with an indolent onset, and its clinical symptoms resemble those of PE. Common manifestations of PAS include exertional dyspnea, chest pain, cough, hemoptysis, fatigue, fever, anemia, weight loss, increased erythrocyte sedimentation rate, and absence of hypercoagulability. The level of D-dimer elevates in patients with PE whereas, in the patients with PAS is usually within the normal range[2-4].

As recent developments, large language models (LLMs) can be used in many areas due to their ability to deeply analyze natural language context, create human-like, consistent, and fluent texts, demonstrate content awareness in information-intensive contexts, and contribute to decision-making and problem-solving processes by understanding instructions[5,6].

This study aimed to evaluate the potential of LLMs in medical decision support processes and, in particular, to comparatively analyze the preliminary and final diagnosis performances of different models and to investigate the ability to make differential diagnosis over PE and PAS cases included using literature data. 18 real medical cases published in the literature were selected[7-18], and for each case, a preliminary diagnosis was obtained from the models by providing only basic clinical information. Then, final diagnoses were collected with an extended dataset including advanced imaging results, laboratory data, and treatment information. Three different platforms were used as LLMs: ChatGPT-4o (OpenAI), Gemini Flash 2.5 (Google DeepMind), and DeepSeek V3 (DeepSeek-AI). Each model was evaluated using fixed and identically structured prompts.

MATERIALS AND METHODS

Study Population

This study reviewed the National Institutes of Health library (https://pubmed.ncbi.nlm.nih.gov) for literature on PE and PAS from 1999 to 2024. Publications that comprised cases with confirmed diagnoses by CT, PET/CT, and histology were chosen at random and included in the evaluation. This study uses data from current articles to perform screening and research utilizing LLM.

Inclusion and Exclusion Criteria

Cases were included if they met the following criteria: (1) the diagnosis of either PE or PAS was confirmed by imaging (e.g., CT angiography, PET-CT) and/or histopathological examination; (2) sufficient clinical details (such as symptoms, D-dimer levels, imaging results, and treatment approach) were available in the case description to construct the diagnostic prompts; (3) the case was published in a peer-reviewed journal between 1999 and 2024. Nine patients presented as a case series in an article were included in the study[19].

Exclusion Criteria

(1) Review articles or editorials without detailed case data, (2) duplicate cases published in more than one article, and (3) cases lacking either a confirmed final diagnosis or the clinical/imaging information required for the two-stage prompt design.

Study Protocols

The study comprised publications that identified the type of article, the age and gender of the patients, and the country to which the item belonged. The D-dimer level as a laboratory test, the diagnostic examinations conducted (echocardiography, Doppler US, CT angiography, MRA, PET/CT), the duration of the disease, the presence of accompanying conditions, the reason for the complaint at the time of application, the onset of the symptoms, and the initial preliminary diagnosis of the patients and the final diagnosis as a result of the techniques performed were all categorized. According to the article data, the location of the PE observed in the CTA and the following treatment strategy (medical, interventional, or surgical) were assessed, as were the PAS size, type, localization, extension, presence of recurrence, and outcome (day) found in these individuals.

Large Language Models

A two-stage evaluation process was adopted for each case included in the study. In the first stage, only the basic clinical

history and initial evaluation findings of the patient were presented to the model to obtain a preliminary diagnosis. In the second stage, detailed information such as advanced imaging results, laboratory data, and treatment process for the same case was presented to the model to obtain a final diagnosis. Thanks to this two-stage structure, the models' diagnostic thinking levels, the way they use information, and their ability to produce meaning in a clinical context were comparatively analyzed.

Three different LLM (ChatGPT-4o, Gemini Flash 2.5, and DeepSeek V3) were evaluated separately for each case using fixed prompt structures. Model outputs were collected manually, and each diagnosis was compared with the final diagnosis of the relevant case stated in the literature.

Prompt Design

The quality of interactions with the LLM is directly dependent on the structure of the prompts used. Prompts are instructions given to the LLM to perform a task. By creating effective prompts, researchers and practitioners can increase the accuracy and relevance of the responses the model provides[6]. In this study, a prompt was created and used with the Role-Goal-Context (RGC) framework.

The RGC framework offers a structured methodology that stands out in terms of clarity and direction. This framework is based on a clear definition of the role the model should adopt, the goal it is expected to achieve, and the contextual information that affects this process. This makes it easier for the model to focus on the task and produce more meaningful, context-sensitive outputs[20,21]. This framework suggests that a prompt should consist of three basic elements:

Role: Specifies the area of expertise in which the model is expected to act.

Goal: Clearly states the specific task the model is expected to answer.

Context: Provides the case information and guiding context needed for the model to produce the answer.

The RGC structure is recommended to increase both consistency and task focus, especially in medical LLM applications. In this context, the prompt structure used in the study was designed in two stages. For each case, first the "preliminary diagnosis prompt", which includes only basic clinical information, and then the "final diagnosis prompt", which includes advanced examinations and treatment processes, was used. Both prompts were prepared to preserve the same RGC structure. The model was asked to produce only a single-sentence, non-explanatory diagnostic response. The prompt used for preliminary diagnosis is given below.

Because no exemplars, demonstrations, or labelled example cases were provided in the prompts, the prompting approach corresponds to a zero-shot setup. This choice was made to *(i)* keep the prompting conditions identical across models, *(ii)* better reflect a first-pass clinical decision support use case, and *(iii)* reduce the risk of implicit guidance from in-prompt examples.

"You are a pulmonary embolism specialist and have in-depth knowledge in the diagnosis and treatment of this disease. You are asked to analyze the following case and make a diagnosis. The following case includes a patient's clinical history and initial evaluation results. Based on these data, make a single sentence for the patient's probable preliminary diagnosis. Make a short sentence for the preliminary diagnosis without explanation." Based on this prompt, the following explanation can be made for the RGC frame.

Role: *You are a pulmonary embolism specialist and have in-depth knowledge of the diagnosis and treatment of this disease.*

Objective: *Analyze the following case and make a diagnosis.*

Context: *The following case includes a patient's clinical history and initial evaluation results. Based on this data, make a single sentence for the patient's probable preliminary diagnosis. Make a short sentence for the preliminary diagnosis without explanation.*

The final diagnosis prompt was created as follows: *"After adding advanced imaging, laboratory, and treatment information about the same case, in light of this information, state the patient's final diagnosis in a single sentence. Make a final diagnosis with a short sentence without explanation."*

This framework allowed for the objectification of the evaluation process and the standard comparison of the responses of various models to the diagnosis procedure. Furthermore, maintaining the prompt language was intended to lessen the impact of bias in model comparisons. For every situation, a different workspace was made.

Assessment and Analysis

The evaluations were made by an experienced radiologist (13 years of clinical experience), and each diagnostic result was classified into three categories:

- If the preliminary diagnosis obtained from LLM according to the given prompts (PE and thrombotic conditions), and the final diagnosis is PAS, labeled as "Exact Match";
- If the preliminary diagnosis in the article for LLM is preliminary diagnosis (PE and thrombotic conditions), and the final diagnosis is PE and thrombotic conditions, labeled as "Approximate Match";
- If the LMM diagnosis and the final diagnosis are unrelated, labeled as "No Match".

The obtained data were systematically tabulated, and the diagnostic accuracy rates were statistically analyzed. This approach provided the opportunity to objectively evaluate the reliability and applicability of LLMs in the context of clinical decision support systems.

The workflow visualizing the process steps and evaluation process followed in the study is presented (Figure 1).

Used LLMs and Access Method

In this study, model outputs were comparatively evaluated using three different LLMs (ChatGPT-4o, Gemini Flash 2.5, and DeepSeek V3) in the diagnosis processes for PE cases. The models were tested directly through user interfaces via open-access platforms; no external intervention, parameter adjustment, or special fine-tuning process was applied. With this method, the raw performance of the models was measured in a natural interaction environment.

The models were accessed via their official web user interfaces not via an application programming interface (API). Because web interfaces may not expose an exact API model identifier and may be updated over time, we report the model labels shown in the interface and the access dates to support reproducibility: ChatGPT-4o [(OpenAI; accessed on 06.07.2025)], Gemini Flash (Google; interface label shown as "Gemini Flash 2.5" at time of access on [(06.07.2025)], and DeepSeek V3 (DeepSeek; accessed on [(07.07.2025)]. Each case was evaluated in a new, empty session to prevent cross-case context leakage.

Statistical Analysis

A statistical package program (SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.) was used to analyze the data obtained from the literature. Descriptive analysis was used for demographic data, chi-square test for categorical data; independent sample T-test was used for normally distributed data in comparisons of binary groups, and Mann-Whitney U test was used for non-normally distributed data. Statistical significance level was accepted as $p<0.05$.

Ethical Considerations

Since this study utilized publicly available case data extracted from previously published literature and did not involve any direct patient contact or identifiable personal information, ethics committee approval was not required in accordance with institutional and international research guidelines.

RESULTS

Of the scanned articles, 77.8% (n=14) were case reports, 22.2% (n=4) were original articles. The most cases were reported in 2024 (n=6), 2018 (n=5).

Female gender (n=10, 55.6%) was the most common. The mean age of the patients was 49.94±13.67 years, and PAS was most frequently detected in the 40-49 and 50-59 age groups (n=5, 27.8%).

The country reporting PAS the most was China (n=11, 61.1%), followed by Türkiye (n=2, 11.1%) and Iran, Brazil, Taiwan, Germany, and the USA (n=1, 5.6%) with equal frequency.

The mean duration of complaints was 67.39±137,85 days, and the mean outcome was 173,89±293,34 days. 38.9% (n=7) of the cases had other accompanying diseases, the most common being post-COVID status (n=2, 11.1%), followed by ARDS, diabetes mellitus, hypertension, pneumonia, and TB (n=1, 5.6%).

The mean D-dimer level in the patients was 596.06±875.78 mg/L 11.1% (n=2) of the cases had a history of anticoagulant drug use, and DMA heparin was used.

Of the patients, 77.8% (n=14) had dyspnea, 16.7% (n=3) syncope, fever, and cough; 11.1% (n=2) experienced palpitations and exhaustion; 44.4% (n=8) experienced chest discomfort; and 27.8% (n=5) experienced hemoptysis. Of those 5.6% (n=1) experienced symptoms that worsened with exercise, the previous day, and the past two hours, while 11.1% (n=2) experienced symptoms that worsened throughout the previous two days.

Echocardiography was performed on 88.9% (n=16) of the patients, Dopple US on 22.2% (n=4), pulmonary CT angiography on 100% (n=18), magnetic resonance imaging (MRI) on 16.7% (n=3), PET/CT on 22.2% (n=4), and MRI on 100% (n=18).

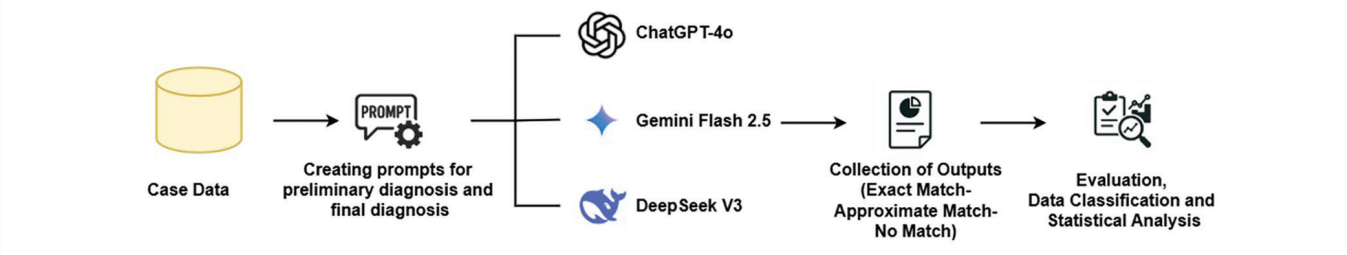

Figure 1. Workflow diagram showing the steps of the operations carried out within the scope of the study

The detected PE localization was mostly at the trifurcation level in 50% (n=9), main-right pulmonary artery level in 16.7% (n=3), main pulmonary artery level in 11.1% (n=2), and right main pulmonary artery level in 5.6% (n=1).

Among the histopathological types of PAS, intimal-mural was the most common (66.7%, n=12), while leiomyosarcoma was the second most common histopathological diagnosis in 11.1% (n=2). Histopathologically, the most common PAS localization was at the trifurcation (61.1%, n=11), main-right pulmonary artery (16.7%, n=3), and main pulmonary artery and right pulmonary artery (11.1%, n=2) were other regions. The intimal-mural type was mostly observed in females (n=7) (Table 1). PAS mostly extended to the pulmonary valve (11.1%, n=2), while aorticopulmonary valve extension to the aorta and right ventricle was reported equally frequently (5.6%, n=1) (Table 1).

The mean three-dimensional (anteroposterior, transverse, craniocaudal) diameters of the PAS in the articles included in the study (n=4, 22.22%) were 3.48±1.86 x 6.18±4.07 x 7.63±8.34 mm. Dimension information was not available in 13 of the articles.

The most preferred method in terms of PE-PAS treatment approach was 38.9% (n=7) surgery, 22.2% (n=4) medical-surgical, 11.1% (n=2) interventional, and 5.6% (n=1) interventional-surgical. The approach was not specified in four cases. Among interventional and surgical methods, the most common approach was surgical (27.8%, n=5); pulmonary endarterectomy, biopsy, right pneumectomy, transcatheter thrombolysis, and inferior vena cava filter, thrombectomy (5.6%, n=1) were other approaches.

The most likely diagnosis for both genders was PE (72.2%, n=13); however, only one male case (n=1) considered the possibility of PAS as a preliminary diagnosis (Table 2).

Comparing literature and LLM data for the preliminary diagnosis of PE, DeepSeek V3 (n=16, 88.89%), ChatGPT-4o (n=15, 83.33%), and Gemini Flash 2.5 (n=12, 66.67%) were the most successful LLM approaches (see Figures 2,3; Table 3).

Table 1. Distribution of PAS histopathological types according to gender

Histopathological type	Gender	
	Male (n,%)	Female (n,%)
Intimal-mural	5 (62.5)	7 (70.0)
Luminal	0	1 (10.0)
Undifferentiated sarcoma	0	1 (10.0)
Leiomyosarcoma	1 (12.5)	1 (10.0)
Rhabdomyosarcoma	1 (12.5)	0
Renal clear cell high-grade sarcoma	1 (12.5)	0
Total	8 (100)	10 (100)

PAS: Pulmonary artery sarcoma

Table 2. Distribution of preliminary diagnoses by gender

Preliminary diagnosis	Gender	
	Male (n,%)	Female (n,%)
PTE	5 (62.5)	8 (80.0)
Pulmonary stenosis	1 (12.5)	0
Chronic thromboembolic pulmonary hypertension	1 (12.5)	0
Pulmonary artery sarcoma	1 (12.5)	0
Pulmonary artery aneurysm	0	1 (10.0)
Malignancy and PTE	0	1 (10.0)
Total	8 (100)	10 (100)

PTE: Pulmonary thromboembolism

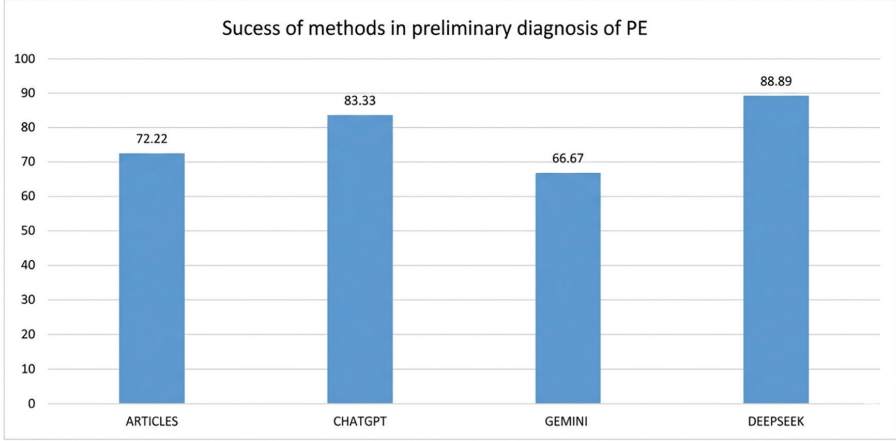

Figure 2. Success distribution of literature and LLM methods in the preliminary diagnosis of PE
LLM: Large language models, PE: Pulmonary embolise

The most successful LLM method for the preliminary diagnosis of PE and the final diagnosis of PAS as a result of the provided case information was Gemini Flash 2.5 (n=4, 22.22%), ChatGPT-4o (n=3, 16.67%), and DeepSeek V3 (n=2, 11.11%) (Figure 4).

DISCUSSION

PAS lesions are continuous, with rounded, bulged, or lobulated surfaces that protrude in the direction of blood flow. They form as tumor tissue accumulates, grows, and invades the surrounding tissues. Regarding filling deficiencies in the pulmonary trunk and pulmonary arteries, the computed tomography pulmonary angiography (CTPA) results for PE and PAS are comparable; nevertheless, the characteristics of each are different. When seen against the blood flow in CT scans, PE shows as cup-like formations. This may be the result of friction caused by blood flow at the surface clot, which is undergoing dissolution by the fibrinolytic system of the blood. Necrosis and bleeding may also occur in PAS, and contrast-enhanced CT can significantly enhance the signals from the blood vessels supplying the PAS tumor, which originates from the pulmonary arteries involved. As a result, the tumor signals in CTPA images are markedly intensified. In contrast, the emboli in patients with PE appear as filling defects with relatively uniform intensities[3]. In this study, the highest PAS detection success of LLMs was observed in Gemini Flash 2.5 (n=4/18, 22.22%).

The World Health Organization classifies PAS into 2 types: wall sarcoma, which is mainly smooth muscle sarcoma, and intimal sarcoma. The most common pathological type of PAS is undifferentiated sarcoma (34%), followed by fibrosarcoma (21%), smooth muscle sarcoma (20%), rhabdomyosarcoma (6%), mesenchymal histiocytoma (6%), intrachondral sarcoma (4%), angiosarcoma (4%), osteosarcoma (3%), and malignant fibrous histiocytoma (2%)[4]. Consistent with the literature, in the current study, the most frequently intimal-mural types were observed.

In this study, Gemini Flash 2.5 had the highest success rate for PAS final diagnosis, and DeepSeek V3 had the highest success rate for PE preliminary diagnosis based on the results of LLM approaches employing literature data. However, when prompts and disease information were supplied, no LMM technique was able to identify PAS as the initial tentative diagnosis. PE was taken into consideration initially, then infectious causes, and chronic thromboembolic pulmonary hypertension. On the other hand, based on the evidence surrounding the condition, there is research that suggests PAS as the initial tentative diagnosis[22]. This shows that although learning techniques, such as deep LLM, are thought to provide support and help in establishing the diagnosis in terms of time and cost, the importance of the human factor in diagnosis and the need for clinical management.

A common characteristic of the cases where model errors were concentrated was that they included individuals who resembled PE but did not have PET/CT (except for cases where PAS was suspected) and whose D-dimer levels were within normal ranges. This implies that LLMs show bias when faced with incomplete data and have poor clinical prediction in uncommon circumstances. Furthermore, notable variations in diagnostic success were noted between the models employed; ChatGPT-4o yielded more accurate results, particularly in cases with limited clinical findings and inadequate imaging support, whereas DeepSeek V3 and Gemini Flash 2.5 models demonstrated lower accuracy rates in comparable cases.

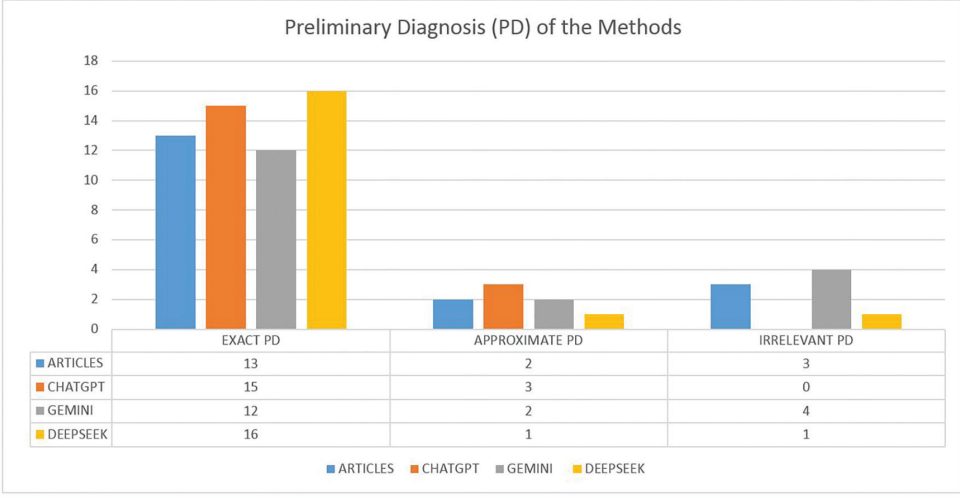

Figure 3. Distribution of LLM methods in the detection of PE preliminary diagnosis
LLM: Large language models, PE: Pulmonary embolise

Table 3. Matching and distribution of LLM methods in preliminary and final diagnosis

	CASE	CHATGPT (preliminary diagnosis)	Final diagnosis	Matching
ChatGPT-4o	CASE 1[2]	Pulmonary embolism	Pulmonary artery sarcoma	EXACT MATCH
	CASE 2[18]	Pulmonary embolism	Acute massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 3[19]	Subsegmental pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 4[19]	Subacute or chronic pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 5[19]	Pulmonary embolism due to chronic thromboembolic pulmonary hypertension	Pulmonary embolism complicated by thrombosis of the main pulmonary artery and right pulmonary artery due to chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 6[19]	Pulmonary embolism	Massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 7[19]	Chronic thromboembolic pulmonary hypertension	Bilateral pulmonary thromboembolism complicated by chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 8[22]	Pulmonary embolism	Massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 9[7]	Pulmonary embolism	Massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 10[17]	Pulmonary embolism compatible with bilateral pulmonary artery embolism	Bilateral massive pulmonary embolism complicated by chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 11[4]	Chronic thromboembolic pulmonary hypertension	Pulmonary artery sarcoma	EXACT MATCH
	CASE 12[9]	Pulmonary embolism	Pulmonary artery sarcoma	EXACT MATCH
	CASE 13[10]	Pulmonary embolism	Pulmonary arterial hemorrhage due to right interlobar pulmonary artery pseudoaneurysm	NO MATCH
	CASE 14[11]	Chronic pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 15[13]	Acute pulmonary embolism	Submassive acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 16[12]	Pulmonary embolism	Central pulmonary embolism	APPROXIMATE MATCHING
	CASE 17[14]	Pulmonary hypertension assive pulmonary embolism compatible with a large filling defect in the main pulmonary artery	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 18[16]	Recurrent or persistent pulmonary thromboembolism.	Chronic thromboembolic pulmonary embolism causing pulmonary hypertension	APPROXIMATE MATCHING

Table 3. Continued

	CASE	Gemini (preliminary diagnosis)	Final diagnosis	Matching
Gemini Flash 2.5	CASE 1[2]	Pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 2[18]	Pulmonary hypertension	Acute proximal pulmonary embolism	APPROXIMATE MATCHING
	CASE 3[19]	Pulmonary hypertension	Pulmoner arteriyel hipertansiyon	NO MATCH
	CASE 4[19]	Acute or chronic pulmonary embolism	Pulmonary artery sarcoma	EXACT MATCH
	CASE 5[19]	Septic pulmonary embolism due to chronic thromboembolic pulmonary hypertension or infective endocarditis	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 6[19]	Acute massive pulmonary embolism	Acute massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 7[19]	Pulmonary hypertension	Chronic thromboembolic pulmonary hypertension	NO MATCH
	CASE 8[22]	Pulmonary tuberculosis	Chronic pulmonary embolism resulting from pulmonary tuberculosis and showing findings compatible with lung infection	APPROXIMATE MATCHING
	CASE 9[7]	Pulmonary embolism	Acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 10[17]	Pulmonary embolism with chronic dyspnea and acute deterioration after COVID-19	Pulmonary artery embolism	APPROXIMATE MATCHING
	CASE 11[4]	Pulmonary embolism	Atypical pulmonary embolism or primary pulmonary artery sarcoma	EXACT MATCH
	CASE 12[9]	Pulmonary embolism or cardiopulmonary pathology	Atypical pulmonary embolism or primary pulmonary artery sarcoma	EXACT MATCH
	CASE 13[10]	Pulmonary embolism	Saccular dilatation of the right pulmonary artery without significant intraluminal filling defect or obstruction, accompanied by moderate right pleural effusion, is an atypical pulmonary artery pathology.	NO MATCH
	CASE 14[11]	Pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 15[13]	Acute pulmonary embolism	Acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 16[12]	Pulmonary embolism	Acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 17[14]	Chronic thromboembolic pulmonary hypertension	Pulmonary artery sarcoma (PAS)	EXACT MATCH
	CASE 18[16]	Chronic thromboembolic pulmonary hypertension or recurrent pulmonary embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING

Table 3. Continued

	CASE	DeepSeek (preliminary diagnosis)	Final diagnosis	Matching
DeepSeek V3	CASE 1[2]	Pulmonary embolism	Pulmonary artery sarcoma (PAS)	EXACT MATCH
	CASE 2[18]	Acute pulmonary embolism	Massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 3[19]	Pulmonary embolism	Acute submassive pulmonary embolism	APPROXIMATE MATCHING
	CASE 4[19]	Pulmonary embolism	Chronic pulmonary embolism	APPROXIMATE MATCHING
	CASE 5[19]	Pulmonary embolism	pulmonary embolism	APPROXIMATE MATCHING
	CASE 6[19]	Acute pulmonary embolism	Acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 7[19]	Massive pulmonary embolism	Massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 8[22]	Pulmonary tuberculosis	Pulmonary tuberculosis	NO MATCH
	CASE 9[7]	Acute pulmonary embolism	Massive bilateral pulmonary embolism	APPROXIMATE MATCHING
	CASE 10[17]	Bilateral multiple pulmonary embolism	Acute right heart failure and pulmonary hypertension due to bilateral multiple pulmonary embolism	APPROXIMATE MATCHING
	CASE 11[4]	Chronic thromboembolic pulmonary hypertension	Pulmonary artery sarcoma (PAS)	EXACT MATCH
	CASE 12[9]	Pulmonary Embolism	Acute massive pulmonary embolism	APPROXIMATE MATCHING
	CASE 13[10]	Acute Pulmonary Embolism	Right pulmonary artery aneurysm	NO MATCH
	CASE 14[11]	Pulmonary Embolism	Chronic thromboembolic pulmonary hypertension	NO MATCH
	CASE 15[13]	Acute Pulmonary Embolism	Acute pulmonary embolism	APPROXIMATE MATCHING
	CASE 16[12]	Acute Pulmonary Embolism	Possible pulmonary embolism	APPROXIMATE MATCHING
	CASE 17[14]	Pulmonary Embolism	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING
	CASE 18[16]	Pulmonary embolism refractory to anticoagulant therapy	Chronic thromboembolic pulmonary hypertension	APPROXIMATE MATCHING

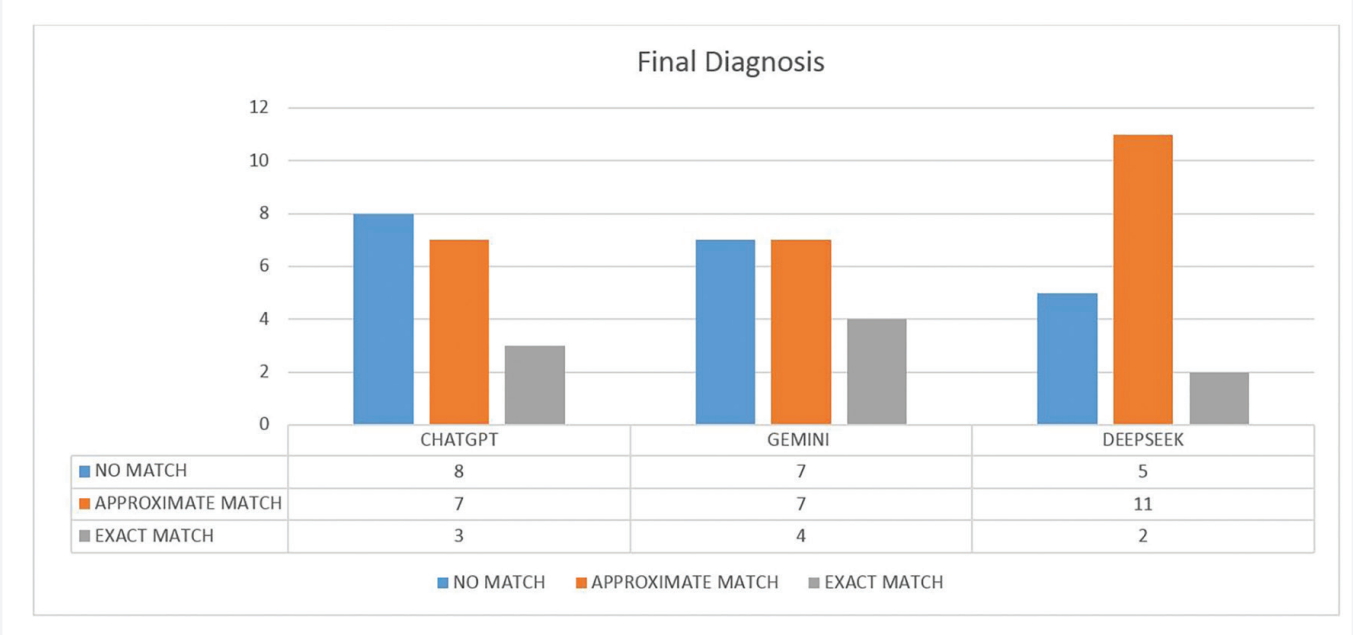

Figure 4. Final diagnosis (PAS) distribution in LLM methods
PAS: Pulmonary artery sarcoma, LLM: Large language models

Research in the literature makes a solid argument for medical practitioners using LLMs to assist clinical practice. Due to the possibility of patient injury, this use of LLMs also entails serious ethical concerns. It may become ethically required to use these technologies in clinical practice if these hazards are reduced by sufficient and trustworthy quality control procedures[5]. However, before this process, data processing methods must be transparent and auditable, and decision-making mechanisms must be structured in line with the principle of responsibility[23].

Despite their outstanding ability to generate responses based on vast amounts of data, LLMs have serious inherent limits when it comes to clinical decision-making. These limitations include a propensity to draw conclusions that defy medical common sense, poor performance in open-ended and ambiguous clinical scenarios, a high degree of confidence in their responses that does not always translate into accuracy, and a failure to perform well on rare conditions or those that are not sufficiently represented in the training data. These issues imply that LLMs should be applied cautiously and moderately in actual clinical settings, as they are caused by the models' rigid thinking and excessive dependence on previously observed patterns[24]. Furthermore, free and open-source models have greater performance limitations than premium and closed-source models, which often have better accuracy rates. As such, model selection should be carefully considered based on the context of use. It should be noted that LLMs are periodically updated by their developers, which may change their diagnostic behavior and output consistency over time. Therefore, reproducibility of results between different time points or model versions may not be guaranteed. LLMs could be used as auxiliary tools in clinical decision support systems, but only with the advice of an expert.

Technical proficiency alone is insufficient to integrate LLMs into clinical applications; a multidisciplinary strategy encompassing ethical principles, legal responsibilities, and regulatory approval procedures is also necessary. Legal liability may emerge when LLM-based decision support systems are used, especially in the healthcare industry, to safeguard patient privacy and consent rights. It is also well known that certain models incorporate user input and cues to enhance the model itself, which raises further moral and legal concerns around patient data privacy.

In a similar study, potential high-risk factors and precautions in cancer were investigated using literature data and LLM. 59 articles were included in the review and were categorized as quantitative studies on LLMs, chatbot-focused studies, and qualitative discussions on LLMs on cancer. Quantitative studies emphasize the advanced capabilities of LLMs in the field of natural language processing, while chatbot-focused articles reveal their potential in clinical support and data management. Qualitative research emphasizes the broader impacts of LLMs, including risks and ethical issues. As the results of the study, Quantitative studies suggest that LLMs may contribute to advancements in diagnostics and patient care, while chatbot-focused studies, particularly on ChatGPT-4o, indicate their potential utility in clinical support and patient communication. Conversely, qualitative analyses reveal concerns about ethics, data privacy, and the need for tailored models. The integration of LLMs in cancer research and healthcare presents a promising avenue for improving patient care[23].

Although LLMs provide valuable contributions in specific clinical contexts, their effectiveness may be restricted in rare diseases or cases with missing data. This circumstance demonstrates that artificial intelligence-only solutions are insufficient for decision support procedures. As a result, according to a comprehensive survey done across Europe, independent experts continue to obtain the greatest accuracy rates. However, models like GPT-4 can yield impressive outcomes, particularly in certain domains like organ estimates. In the study conducted by Saban et al.[25], it was emphasized that hybrid approaches using human expertise and artificial intelligence together can be one of the most effective methods in justifying imaging requests. Our study similarly demonstrates that LLMs should only be considered as supportive tools in clinical decision support processes and that final decisions should be made under expert supervision. Hybrid human-AI approaches, where LLMs are used together with expert supervision, are considered to be the most appropriate strategy for increasing diagnostic accuracy and safely optimizing clinical decision support processes.

Study Limitations

Data leakage is a potential limitation of our design because all evaluated cases were extracted from previously published literature. Therefore, some case narratives or key phrases may have appeared in the training corpora of the evaluated LLMs, and model outputs could partly reflect memorisation rather than clinical reasoning. To mitigate this risk, we provided only structured clinical summaries without author names, journal titles, or direct quotations, and we did not use any retrieval or browsing features. Nevertheless, the possibility of partial exposure cannot be fully excluded; thus, our results should be interpreted as reflecting performance in a real-world setting where models may have prior exposure to published case reports. Future work should include unpublished or prospectively collected cases, and/or synthetic cases with altered details to more rigorously quantify memorisation effects.

Additionally, the relatively small number of cases (n=18) and the single-run evaluation per case should be considered exploratory; future studies with larger, prospectively collected datasets and repeated runs per model to assess output stability are warranted.

CONCLUSION

The diagnoses given by the models were compared with human expert opinions and case results in the literature, and their potential in terms of accuracy, clinical consistency, and decision support systems was interpreted. In this context, the study presents an interdisciplinary evaluation questioning the usability of LLMs in diagnostic processes based on structured information in the field of health. Although artificial intelligence systems and LLM can be used to evaluate and improve the effectiveness of the diagnosis and treatment process, PAS should be considered in the differential diagnosis of PE cases that do not improve despite treatment, in addition to the patient's clinical history and current laboratory data.

Ethics

Ethical Committee Approval: Since this study utilized publicly available case data extracted from previously published literature and did not involve any direct patient contact or identifiable personal information, ethics committee approval was not required in accordance with institutional and international research guidelines.

Declaration on the use of Artificial Intelligence (AI): During the preparation of this work the author(s) used QuillBot AI in order to improve the readability and language of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Footnotes

Authorship Contributions

Concept: H.S., Design: H.S., M.A.Ş., Data Collection or Processing: H.S., Analysis or Interpretation: H.S., M.A.Ş., Literature Search: H.S., M.A.Ş., Writing: H.S., M.A.Ş.

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

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

REFERENCES

1. Moore AJE, Wachsmann J, Chamarthy MR, Panjikaran L, Tanabe Y, Rajiah P. Imaging of acute pulmonary embolism: an update. Cardiovasc Diagn Ther. 2018;8:225-43.
2. Liu Z, Fan L, Liang S, Wu Z, Huang H. A primary pulmonary artery sarcoma masquerading pulmonary embolism: a case report and literature review. Thromb J. 2024;22:4.
3. Pu X, Song M, Huang X, Zhu G, Chen D, Gan H, et al. Clinical and radiological features of pulmonary artery sarcoma: a report of nine cases. Clin Respir J. 2018;12:1820-9.
4. Wang Y, Rong C, Liu J, Liu X, Zhang W. Pulmonary arterial sarcoma: a case report. Medicine (Baltimore). 2024;103:e37194.
5. Armitage RC. How do GPs want large language models to be applied in primary care, and what are their concerns? A cross-sectional survey. J Eval Clin Pract. 2025;31:e70129.
6. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res. 2024;26:e60501.
7. Shen C, Xu W, Ouyang R. Primary pulmonary artery sarcoma complicated with pulmonary embolism and pulmonary tuberculosis: a case report and literature review. 2022;47:673-8.
8. Jin T, Zhang C, Feng Z, Ni Y. Primary pulmonary artery sarcoma. Interact Cardiovasc Thorac Surg. 2008;7:722-4.
9. Atahan C, Güral Z, Yücel S, Ağaoğlu F. Pulmonary artery intimal sarcoma: case report of a patient managed with multimodality treatment and a comprehensive literature review. Strahlenther Onkol Organ Dtsch Rontgengesellschaft Al. 2024;200:725-9.
10. Zhao M, Nie P, Guo Y, Chen H. Pulmonary artery intimal sarcoma: a rare cause of filling defects in pulmonary arteries. Am J Med Sci. 2022;364:655-60.
11. Terra RM, Fernandez A, Bammann RH, Junqueira JJM, Capelozzi VL. Pulmonary artery sarcoma mimicking a pulmonary artery aneurysm. Ann Thorac Surg. 2008;86:1354-5.
12. Gao X, Xie A, Xiao W, Wei Z, Yu S. Pulmonary artery sarcoma misdiagnosed as pulmonary embolism. J Cardiothorac Vasc Anesth. 2024;38:2041-6.
13. Chen PW, Liu PY. Pulmonary artery sarcoma mimicking pulmonary embolism. BMJ Case Rep. 2018;2018:bcr2018226999.
14. Dörr A, Flörcken A, Bullinger L, Capper D, Deimling AV, Kaul D, et al. Thrombus or tumor? A case report of a rare sarcoma entity: intimal sarcoma of the pulmonary arteries. Mol Biol Rep. 2024;51:568.
15. Gutiérrez A, Sauler M, Mitchell JM, Siegel MD, Trow TK, Bacchetta M, et al. Unresolved pulmonary embolism leading to a diagnosis of pulmonary artery sarcoma. Heart Lung. 2014;43:574-6.
16. Yazgan C, Ertürk H, Taskin A. Unusual cause of filling defect in pulmonary artery: pulmonary artery sarcoma. Pan Afr Med J. 2020;35:41.
17. Li X, Qi Q, Liang F, Zhang X, Dong S, Song B. Primary pulmonary artery sarcoma with deep vein thrombosis: a case report. Medicine (Baltimore). 2019;98:e15874.
18. Weijmer MC, Kummer JA, Thijs LG. Case report of a patient with an intimal sarcoma of the pulmonary trunk presenting as a pulmonary embolism. Neth J Med. 1999;55:80-3.
19. Pu X, Song M, Huang X, Zhu G, Chen D, Gan H, et al. Clinical and radiological features of pulmonary artery sarcoma: a report of nine cases. Clin Respir J. 2018;12:1820-9.
20. Rosa S. Large language models for requirements engineering. [Master's thesis]. Politecnico di Torino; 2025. Available from: https://webthesis.biblio.polito.it/35574/
21. de Santana VF, Berger S, Machado T, de Macedo MMG, Sanctos CS, Williams L, Wu Z. Can LLMs recommend more responsible prompts? In: proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25). New York: association for computing machinery. 2025:298-313.
22. Rashidi F, Bilehjani E, Mousavi-Aghdas SA, Parvizi R. Massive primary pulmonary artery rhabomyosarcoma: a case report. Rom J Intern Med. 2024;62:67-74.
23. Zitu MM, Le TD, Duong T, Haddadan S, Garcia M, Amorrortu R, et al. Large language models in cancer: potentials, risks, and safeguards. BJR Artif Intell. 2024;2:ubae019.
24. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep. 2025;15:39426.
25. Saban M, Alon Y, Luxenburg O, Singer C, Hierath M, Karoussou Schreiner A, et al. Comparison of CT referral justification using clinical decision support and large language models in a large European cohort. Eur Radiol. 2025;35:6150-9.