Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263696

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Large-scale evaluation of multimodal large language models for pneumothorax detection

Hamza Eren Güzel[1]
Cemre Özenbaş[2]
Ali Murat Koç[3]

[1]University of Health Sciences Türkiye, İzmir City Hospital, Clinic of Radiology, İzmir, Türkiye

[2]Tınaztepe University Private Buca Hospital, Department of Radiology, İzmir, Türkiye

[3]İzmir Katip Çelebi University Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye

PURPOSE

Pneumothorax requires rapid recognition and accurate interpretation of chest X-rays (CXRs), particularly in acute settings where delays can have serious consequences. With the emergence of advanced image interpretation models capable of visual analysis, their diagnostic reliability in radiology practice remains to be determined. This study aimed to assess the diagnostic performance of three state-of-the-art systems in detecting pneumothorax using a large, well-annotated dataset.

METHODS

A total of 10,675 CXRs from the publicly available SIIM-ACR Pneumothorax Segmentation dataset were analyzed. Three multimodal models (GPT-4o, Gemini 2 Pro, and Claude 4 Sonnet) were evaluated using a uniform, image-based approach. Each model's binary outputs (presence: 1, absence: 0) were compared with reference results to determine accuracy, sensitivity, specificity, precision, and F1 scores. Additional subgroup analyses were conducted across pneumothorax size categories: small, medium, and large. Pairwise statistical comparisons were performed using McNemar's test. Sensitivity, specificity, and overall accuracy are reported with corresponding 95% confidence intervals.

RESULTS

The prevalence of pneumothorax in the dataset was 22.3% (n = 2,379). All models demonstrated high specificity (above 0.90) but consistently low sensitivity (0.16–0.36). The best overall performance was observed with Gemini 2, which achieved an accuracy of 0.79 and specificity of 0.95, whereas Claude 4 showed greater sensitivity (0.20–0.34) across lesion-size categories. Diagnostic performance improved with increasing pneumothorax size, but smaller lesions remained difficult to identify. Pairwise comparisons confirmed statistically significant differences among all evaluated systems ($P < 0.050$).

CONCLUSION

In this large-scale evaluation, the tested models exhibited strong reliability in identifying normal examinations but limited ability to detect subtle or small pneumothoraxes. Despite high specificity, low sensitivity limits the use of current Multimodal large language models as rule-out tools for pneumothorax. With continued refinement, these models may eventually support radiologists by improving workflow efficiency and diagnostic confidence.

CLINICAL SIGNIFICANCE

Automated systems capable of high specificity but low sensitivity should not be relied upon to exclude pneumothorax. However, they may serve as valuable assistants for confirming positive findings and prioritizing urgent cases in busy clinical workflows.

KEYWORDS

Pneumothorax, multimodal large language models, chest X-ray, artificial intelligence, diagnostic accuracy

Corresponding author: Hamza Eren Güzel

E-mail: hamzaerenguzel@gmail.com

Received 07 November 2025; revision requested 01 January 2026; accepted 31 January 2026.

Epub: 12.02.2026

DOI: 10.4274/dir.2026.263696

You may cite this article as: Güzel HE, Özenbaş C, Koç AM. Large-scale evaluation of multimodal large language models for pneumothorax detection. *Diagn Interv Radiol.* 12 February 2026 DOI: 10.4274/dir.2026.263696 [Epub Ahead of Print].

Pneumothorax is a potentially life-threatening condition characterized by the presence of air within the pleural space, leading to partial or complete lung collapse. The diagnostic approach to pneumothorax should begin with a standard posteroanterior chest X-ray (CXR), which remains the first-line imaging modality for confirming the presence of intrapleural air and assessing its extent. Recent clinical guidelines highlight that timely and accurate interpretation of CXRs is critical, as management decisions often depend on prompt recognition and precise evaluation of pneumothorax size and severity.[1]

Over the past decade, deep learning (DL) has shown strong performance in the automated detection and segmentation of pneumothorax on CXR. Several systematic reviews and large studies have reported high diagnostic accuracy across different datasets, underscoring the potential role of DL as a triage or second-reader tool.[2-5] These advances were made possible not only by improvements in algorithms but also by the availability of large, open datasets that helped standardize benchmarking in pneumothorax detection.[4]

Multimodal large language models (MLLMs) are general-purpose artificial intelligence (AI) systems capable of processing both visual and textual inputs. Although prior studies have reported promising results when such models are provided with combined image and clinical context, their performance in image-only diagnostic tasks remains inconsistent.[6-10] However, most current MLLMs are not specifically trained for medical imaging tasks, which may limit their spatial localization and diagnostic precision, particularly when relying solely on image-based inputs. In this context, evaluating MLLMs not only on small patient cohorts but also on large, well-curated datasets is crucial to establishing their reliability and diagnostic validity.

Main points

- Multimodal large language models (GPT-4o, Claude 4, Gemini 2) were evaluated on more than 10,000 chest X-rays for pneumothorax detection.
- All models achieved high specificity but low sensitivity, in all sizes.
- They may serve as potential rule-in aids for radiologists, but their low sensitivity makes rule-out use clinically unsafe.

In this study, we aimed to evaluate the performance of MLLMs in detecting pneumothorax on CXRs using a large-scale dataset. In addition, we assessed the ability of different MLLMs to identify pneumothoraxes of varying sizes. We hypothesized that the diagnostic performance of general-purpose MLLMs for pneumothorax detection on CXRs would differ between models and across pneumothorax size categories. Given the critical importance of rapid diagnosis in emergency settings, our study also sought to explore the practical applicability and limitations of these rapidly evolving models in the detection of pneumothorax.

Methods

Dataset

We used the publicly available SIIM-ACR Pneumothorax Segmentation Dataset, originally developed for the 2019 Kaggle challenge.[11] The dataset provides CXRs in DICOM format, annotated with pixel-level segmentation masks for pneumothorax. No exclusion criteria were applied, as the dataset is a publicly available, pre-curated benchmark dataset with standardized annotation and quality control procedures. All DICOM images were successfully processed, and no preprocessing failures or unreadable images were encountered. After preprocessing, a total of 10,675 radiographs were included in the analysis. Each radiograph was classified as pneumothorax-positive or -negative according to the presence or absence of a segmentation mask, derived from the SIIM-ACR radiologist annotations, which served as the reference standard. For positive cases, the lesion area was calculated from the mask and used for size categorization. The overall study workflow, including data preprocessing, model inference via application programming interface (API), and comparative performance evaluation, is summarized in Figure 1.

Image preprocessing

All DICOM images were converted into 8-bit grayscale PNG format using a standardized a Python-based workflow was implemented using the Pydicom library (Pydicom, Boston, MA, USA) and the Pillow imaging library (Python Imaging Library, San Francisco, CA, USA). The pipeline ensured consistent windowing, normalization, and resizing across all studies to maintain compatibility and reproducibility.

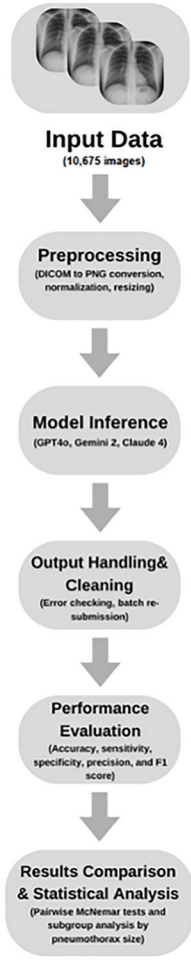

Figure 1. Overview of the study workflow, showing sequential steps from data input and preprocessing to model inference and results comparison.

Each DICOM file was decoded using Pydicom, and the corresponding modality and VOI LUT transformations were applied to preserve true intensity values.

Pixel values were normalized to the 0–255 range through percentile-based clipping (0.5^{th}–99.5^{th}) to reduce the influence of outliers and scanner-specific variations. All images were resized to 1024 × 1024 pixels, preserving the original aspect ratio via zero padding to avoid geometric distortion.

Each image was saved as an 8-bit grayscale PNG file suitable for MLLM inference. A conversion manifest linking each PNG to its source DICOM (SOPInstanceUID) was automatically generated for traceability (Figure 2).

Models and API access

Three MLLMs were evaluated: GPT-4o (OpenAI), Gemini 2 Pro (Google DeepMind), and Claude 4 Sonnet (Anthropic).

Each model was accessed through its official API during October and November 2025. Exact model version strings and API access dates are provided in Supplementary Table 1. All images were submitted as standardized 8-bit PNG files prepared during preprocessing. Models were used as provided by the vendors, without additional fine-tuning or training. To ensure comparability, inference was run with deterministic decoding (temperature: 0) and a capped output budget (maximum 512 tokens). For each case, the model response, image identifier, timestamp, and model name were recorded to comma-separated files and subsequently merged with the master analysis table by filename/ImageId for performance calculations. All inferences were executed via paid API accounts procured by the authors; the vendors had no role in study design, analysis, or manuscript preparation.

Prompting strategy and output handling

A single, model-agnostic prompt was used for all systems to minimize bias. The exact instruction was:

"You are a radiology assistant. Carefully analyze this CXR for pneumothorax.

Return exactly ONE character: 1 if pneumothorax is present or strongly suspected, and 0 if absent.

Answer only with 1 or 0."

The prompt intentionally avoided examples or narrative phrasing to reduce stylistic drift between models and to standardize outputs. Returned texts were parsed with a strict rule set that extracted the first valid numeric character; if the output was non-conforming, defined as any response that did not contain a single numeric character ("1" or "0"), such as textual answers ("Yes," "No," "Present," "Absent"), symbols, or blank strings, the request was repeated once to allow for correction. Persistently unparsable cases (i.e., outputs that remained non-numeric after two iterations) were flagged and subsequently re-submitted to the API as a separate batch. The new responses were manually verified and recorded by the study team to ensure completeness and consistency.

The full Python scripts used for preprocessing, API-based inference, and evaluation are publicly available.[12]

Reference standard and lesion categorization

The ground truth was defined by the segmentation masks provided in the SIIM-ACR dataset. Pneumothorax lesion area was computed in pixels. For pneumothorax-positive cases, lesion size was quantified as the total number of pixels within the segmentation mask after standardization to a 1024 × 1024 grid. Histogram analysis showed a right-skewed distribution: most cases were < 10,000 pixels, although a minority extended above 100,000 pixels. Percentile analysis revealed that the 25th, 50th (median), 75th, 90th, and 95th percentiles corresponded to 4,043, 8,666, 18,732, 34,979, and 48,550 pixels, respectively. Based on this distribution, and to ensure clinically meaningful and statistically balanced subgroups, we defined three categories: small (< 10,000 pixels), medium (10,000–35,000 pixels), and large (> 35,000 pixels). These thresholds were established a priori (before model evaluation) to avoid bias, and were chosen to reflect both the quartile distribution of lesion sizes and the expected clinical conspicuity of different pneumothorax volumes. In cases with multiple separate regions, the total combined mask area was used for categorization.

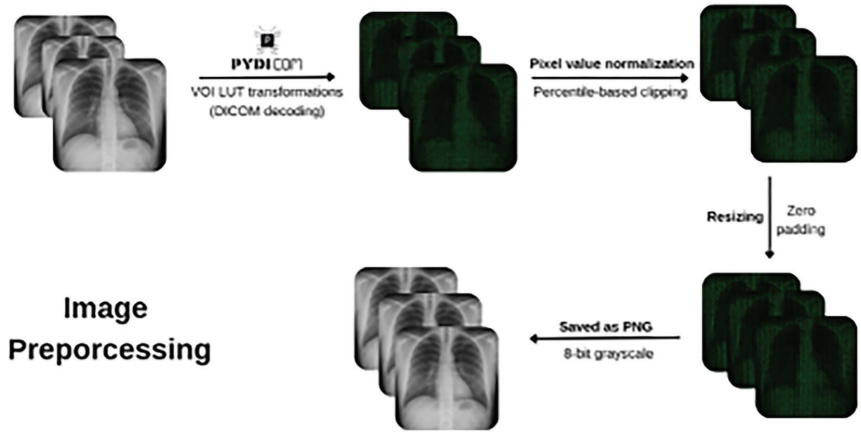

Figure 2. DICOM-to-PNG preprocessing workflow for standardized image conversion.

Statistical analysis

Diagnostic performance was evaluated at the case level using standard classification metrics, including sensitivity, specificity, accuracy, precision, and F1 score. Reporting followed recommended practices for diagnostic accuracy studies.[13] Metrics were calculated for the overall dataset and separately for each lesion size category (small, medium, large). True-positive, false-positive, true-negative, and false-negative counts were derived from model predictions and reference labels. Pairwise comparisons between models were conducted using McNemar's test to assess significant differences in classification performance. A P value < 0.050 was considered statistically significant. All analyses were conducted in Python (version 3.10) using pandas (version 2.2) for data management and scikit-learn (version 1.5) for metric computation.

Ethical considerations

This study was approved by the İzmir City Hospital Ethical Committee (approval number: 2025/550, date: 08.10.2025). As the SIIM-ACR dataset is publicly available and fully de-identified, the requirement for individual informed consent was waived. The study complies with the principles of the Declaration of Helsinki.

Results

Dataset characteristics

A total of 10,675 CXRs were included in the analysis. Of these, 2,379 (22.3%) were labeled as pneumothorax-positive and 8,296 (77.7%) as negative. Among the positive cases, 1,302 (54.7%) were categorized as small, 839 (35.3%) as medium, and 238 (10.0%) as large pneumothoraxes.

Model performance

All three MLLMs were evaluated against the reference standard. Their performance metrics, including accuracy, sensitivity, specificity, precision, and F1 score, are summarized in Table 1.

Although individual results varied across models, all systems achieved high specificity with comparatively lower sensitivity values.

Pairwise comparisons using McNemar's test showed statistically significant differences among the models (P < 0.001). Gemini 2 demonstrated higher overall accuracy and specificity than both Claude 4 and GPT-4o (P < 0.001). Claude 4 exhibited higher sensitivity than Gemini 2 (P = 0.033). GPT-4o showed

significantly lower performance compared with both models across all evaluated metrics ($P < 0.001$). Detailed statistical comparisons are presented in Table 2.

Performance by pneumothorax size

When performance was analyzed by pneumothorax size, metric values differed across subgroups (Table 3).

For small pneumothoraxes (n = 1,302), accuracies ranged from 0.90 (GPT-4o) to 0.93 (Gemini 2), sensitivities from 0.17 to 0.20, and F1 scores from 0.28 to 0.34.

For medium pneumothoraxes (n = 839), accuracies were between 0.92 and 0.94, sensitivities between 0.16 and 0.25, and F1 scores between 0.28 and 0.40.

For large pneumothoraxes (n = 238), accuracies ranged from 0.98 to 0.99, sensitivities from 0.19 to 0.36, and F1 scores from 0.32 to 0.53 across the three models.

Visual representation

The visual overview of model performance highlights how sensitivity increases with pneumothorax size and how true and false classifications are distributed across the dataset, offering an intuitive understanding of the results (Figures 3 and 4).

Table 1. Overall performance of multimodal large language models in pneumothorax detection

Model	Accuracy	Sensitivity	Specificity	Precision	F1 score
GPT-4o	0.75	0.17	0.92	0.36	0.23
Claude 4	0.77	0.23	0.92	0.46	0.31
Gemini 2	0.79	0.22	0.95	0.56	0.31

Table 2. Pairwise McNemar test for differences in model performance

Comparison	Accuracy (*P*)	Sensitivity (*P*)	Specificity (*P*)	Direction of difference
Gemini 2 vs. Claude 4	< 0.001	0.033	< 0.001	Gemini 2: higher accuracy, specificity Claude 4: higher sensitivity
Gemini 2 vs. GPT-4o	< 0.001	< 0.001	< 0.001	Gemini 2: higher in all metrics
Claude 4 vs. GPT-4o	< 0.001	< 0.001	0.21	Claude 4: higher accuracy, sensitivity No difference in specificity

Table 3. Performance stratified by pneumothorax size

Model	Size	Accuracy	Sensitivity	F1 score
GPT-4o	Small	0.90	0.17	0.28
	Medium	0.92	0.16	0.28
	Large	0.98	0.19	0.32
Claude 4	Small	0.91	0.20	0.34
	Medium	0.93	0.25	0.40
	Large	0.98	0.34	0.50
Gemini 2	Small	0.93	0.17	0.29
	Medium	0.94	0.24	0.39
	Large	0.99	0.36	0.53

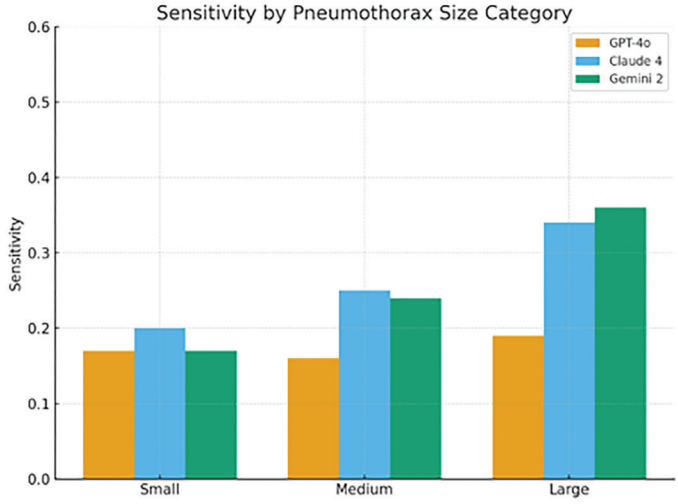

Figure 3. Sensitivity across pneumothorax size categories. All models improved with increasing lesion size, with Gemini 2 consistently achieving the highest sensitivity.

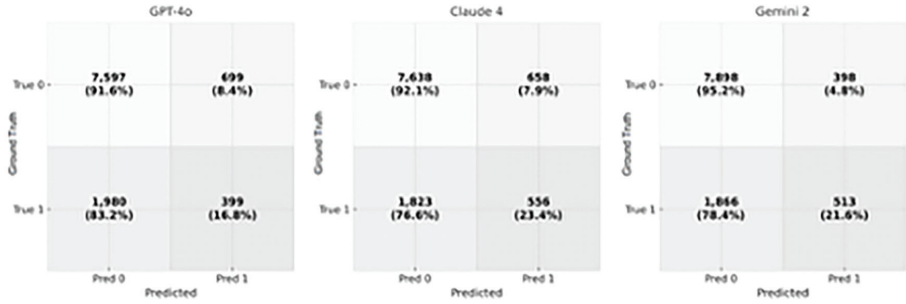

Figure 4. Combined confusion matrices for GPT-4o, Claude 4, and Gemini 2 on the full cohort (n = 10,675; positives: 2,379; negatives: 8,296). Cells display case counts and row-wise percentages within the true class (True 0/True 1); higher diagonal values indicate correct classifications (TN, TP), with axes labeled "True" (rows) and "Predicted" (columns).

TN, true negative; TP, true positive.

Discussion

This large-scale evaluation demonstrated that MLLMs achieve high specificity but consistently low sensitivity for pneumothorax detection on CXRs. This study evaluated three state-of-the-art MLLMs (GPT-4o, Claude 4, and Gemini 2) on the task of pneumothorax detection from CXRs using the SIIM-ACR dataset. The results revealed a consistent pattern: all three models achieved high specificity (> 0.90), whereas sensitivity remained low across all pneumothorax categories, ranging from 0.16 to 0.36. These findings highlight a critical limitation of current MLLMs, raising substantial concerns about their potential use in clinical scenarios such as pneumothorax, where rapid diagnosis and timely intervention are essential.

Numerous AI studies based on imaging data have been conducted to address emergency conditions such as pneumothorax, where clinicians and radiologists must make rapid diagnostic decisions. DL–based approaches, particularly convolutional neural networks, have shown remarkable potential for the automatic detection and localization of pneumothorax on CXRs. For example, Cho et al.[14] investigated a DL–based method for detecting and localizing pneumothorax on CXRs, aiming to enhance diagnostic accuracy and support clinical workflows in emergency settings. They demonstrated that their approach can accurately identify pneumothorax on CXRs, reduce diagnostic delays, and support more effective clinical decision-making and patient care. Similarly, Hillis et al.[15] reported that an AI model could reliably identify both pneumothorax and tension pneumothorax on CXRs. Thian et al.[16] conducted a multicenter external validation study evaluating DL systems for pneumothorax detection on CXRs**.** Their model was tested on datasets from multiple institutions and demonstrated robust generalizability, achieving area under the curve values ranging from 0.91 to 0.98 across external cohorts.

By contrast, there are only a limited number of studies investigating the use of LLMs for pneumothorax detection. Among them, Ostrovsky[17] evaluated the performance of ChatGPT in interpreting CXRs across various thoracic pathologies. In the pneumothorax subgroup consisting of 200 patients, the model achieved a sensitivity of 0.77 and a specificity of 0.89, demonstrating moderate diagnostic capability in this emergency setting. In our study, MLLMs showed noticeably lower diagnostic performance when compared with AI tools developed specifically for image interpretation. Using a large publicly available dataset of more than 10,000 images, our evaluation indicates that current MLLMs still lack the reliability required for routine use in radiology practice.

There are inherent risks associated with integrating AI applications and MLLMs into clinical practice without fully testing their reliability and validity. In this context, the low sensitivity observed in our study could lead to serious problems in clinical practice, particularly in emergency settings. Clinicians relying entirely on these systems for decision-making may delay necessary interventions, potentially worsening the patient's condition. Conversely, the specificity rates exceeding 90% observed in our study suggest that, in their current state, MLLMs may be more suitable for rule-in purposes rather than definitive diagnosis.

In recent years, numerous studies have evaluated the performance of MLLMs in interpreting medical images in the field of radiology. In the study conducted by Sonoda et al.,[8] the diagnostic performance of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro was systematically compared using RSNA Diagnosis Please cases. The results showed that Claude 3 Opus achieved the highest primary diagnostic accuracy at 54%, followed by GPT-4o at 41%, and Gemini 1.5 Pro at 33.9%. Moreover, the proportion of cases where the correct diagnosis was included within the top three differential diagnoses was 62%, 49.4%, and 41%, respectively. Similarly, a study conducted by Suh et al.[18] compared GPT-4V and Gemini Pro Vision using only image inputs against board-certified radiologists. The results demonstrated that GPT-4V reached an accuracy of 49% and Gemini Pro Vision achieved 39%, both falling below the 61% accuracy observed among radiologists. Hirano et al.[19] evaluated multiple MLLMs on the Japanese Diagnostic Radiology Board Examination using image-based questions and found that o3 achieved 72% accuracy, o4-mini and Gemini 2.5 Pro each reached 70%, and Claude 3.7 Sonnet received lower legitimacy scores. Similarly, Nakaura et al.[20] compared several models on board examination items and reported that GPT-4o achieved the highest accuracy (45%), with Claude 3 Opus performing best in text-only tasks (46%).

Despite the varying results across different imaging modalities, one thing is clear: MLLMs are improving rapidly. Hou et al.[21] reported that next-generation models, such as o1 and GPT-4o, demonstrated substantial performance gains compared with their predecessors, with diagnostic accuracy approaching that of human experts in certain scenarios. Although their integration into routine radiology workflows may seem challenging at the moment, it is important to remember that MLLMs have received no training specifically focused on radiologic diagnosis. Given their rapid progress, it seems likely that they may soon perform at a level comparable to dedicated radiology AI systems.

This study has several limitations. First, the evaluated MLLMs were not specifically trained for medical imaging, reflecting a limitation inherent to model design. Second, although these models can process multimodal inputs, the evaluation in this study was limited to image-only prompts, excluding potentially informative clinical or textual context that might enhance diagnostic performance. The use of a simplified binary output format restricted the models' ability to convey diagnostic uncertainty, which may have affected their ability to detect borderline cases. Third, the SIIM-ACR Pneumothorax dataset, although large and well-curated, originates from a single competition source and may not reflect the full diversity of re-

al-world imaging conditions, acquisition parameters, or institutional variations, limiting generalizability. Fourth, the reproducibility of our findings is inherently limited by the use of API-based MLLMs. These systems are subject to ongoing backend updates and undocumented changes implemented by vendors, which may alter model behavior over time, even when identical prompts and inference parameters are used. Lastly, the findings were derived from a single dataset without external validation; future studies using independent or real-world hospital data are required to confirm robustness and clinical applicability.

In this large-scale evaluation using the SIIM-ACR Pneumothorax dataset, MLLMs demonstrated high specificity but low sensitivity in detecting pneumothorax on CXRs. From a clinical perspective, this finding indicates that current MLLMs may serve as rule-in aids for confirming pneumothorax but are unsuitable for rule-out use. Although models such as Gemini 2 and Claude 4 showed relatively better overall performance than GPT-4o, all systems missed the majority of small pneumothoraxes, limiting their clinical reliability in emergency or screening contexts. These findings indicate that current MLLMs, which are not specifically trained for radiologic interpretation, are not yet suitable for standalone diagnostic use. Nevertheless, future iterations or domain-adapted versions of MLLMs may eventually support radiologists by improving workflow efficiency and enhancing diagnostic confidence, particularly as model architectures and training strategies continue to evolve.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table 1. https://d2v96fx-pocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/content-images/93147689-f3e8-4214-b025-72c1b0c25421.pdf

References

1. Roberts ME, Rahman NM, Maskell NA, et al. British Thoracic Society Guideline for pleural disease. *Thorax*. 2023;78(Suppl 3):s1-s42. [Crossref]
2. Sugibayashi T, Walston SL, Matsumoto T, Mitsuyama Y, Miki Y, Ueda D. Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis. *Eur Respir Rev*. 2023;32(168):220259. [Crossref]
3. Thian YL, Ng D, Hallinan JTPD, et al. Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. *Radiol Artif Intell*. 2021;3(4):e200190. [Crossref]
4. Wang Y, Lu Z. Automatic segmentation of pneumothorax in chest radiographs based on dual-task interactive learning method. In: Proceedings of the 2023 5th International Conference on Image, Video and Signal Processing (IVSP '23); 2023 Mar 24-26; Hong Kong, China. New York: ACM; 2023. p. 51-58. [Crossref]
5. Güzel HE, Aşcı G, Demirbilek O, Özdemir TD, Erekli PB. Diagnostic precision of a deep learning algorithm for the classification of non-contrast brain CT reports. *Front Radiol*. 2025;5:1509377. Erratum in: *Front Radiol*. 2025;5:1744006. [Crossref].
6. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-V4 (GPT-4 with Vision) on detection of radiologic findings on chest radiographs. *Radiology*. 2024;311(2):e233270. Erratum in: *Radiology*. 2024;311(2):e249016. [Crossref]
7. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology*. 2024;312(3):e240153. [Crossref]
8. Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J *Radiol*. 2024;42(11):1231-1235. [Crossref]
9. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(12):1399-1402. [Crossref]
10. Büyüktoka RE, Salbas A. Multimodal large language models for pediatric bone-age assessment: a comparative accuracy analysis. *Acad Radiol*. 2025;32(11):6905-6912. [Crossref]
11. SIIM ACR Pneumothorax Segmentation Data [Internet]. Kaggle. [Crossref]
12. MLLM Pneumothorax Detection on SIIM-ACR Dataset [Internet]. GitHub. [Crossref]
13. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283-3289. [Crossref]
14. Cho Y, Kim JS, Lim TH, Lee I, Choi J. Detection of the location of pneumothorax in chest X-rays using small artificial neural networks and a simple training process. *Sci Rep*. 2021;11(1):13054. [Crossref]
15. Hillis JM, Bizzo BC, Mercaldo S, et al. Evaluation of an artificial intelligence model for detection of pneumothorax and tension pneumothorax in chest radiographs. *JAMA Netw Open*. 2022;5(12):e2247172. [Crossref]
16. Thian YL, Ng D, Hallinan JTPD, et al. Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. *Radiol Artif Intell*. 2021;3(4):e200190. [Crossref]
17. 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. [Crossref]
18. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology*. 2024;312(1):e240273. [Crossref]
19. Hirano Y, Miki S, Yamagishi Y, et al. Assessing accuracy and legitimacy of multimodal large language models on Japan Diagnostic Radiology Board Examination. *Jpn J Radiol*. 2026;44(1):209-217. [Crossref]
20. Nakaura T, Yoshida N, Kobayashi N, et al. Performance of multimodal large language models in Japanese Diagnostic Radiology Board Examinations (2021-2023). *Acad Radiol*. 2025;32(5):2394-2401. [Crossref]
21. Hou B, Mukherjee P, Batheja V, Wang KC, Summers RM, Lu Z. One year on: assessing progress of multimodal large language model performance on RSNA 2024 Case of the Day Questions. *Radiology*. 2025;316(2):e250617. [Crossref]