



# Artificial intelligence in radiology: diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans

✉ Olga Bayar-Kapıcı<sup>1</sup>  
 ✉ Erman Altunışık<sup>2</sup>  
 ✉ Feyza Musabeyoğlu<sup>2</sup>  
 ✉ Şeyda Dev<sup>2</sup>  
 ✉ Ömer Kaya<sup>3</sup>

<sup>1</sup>Seyhan State Hospital, Clinic of Radiology, Adana, Türkiye

<sup>2</sup>University of Health Sciences Türkiye, Gaziantep City Hospital, Clinic of Neurology, Gaziantep, Türkiye

<sup>3</sup>Çukurova University Faculty of Medicine, Department of Radiology, Adana, Türkiye

## PURPOSE

Chat Generative Pre-trained Transformer (ChatGPT)-4V, a large language model developed by OpenAI, has been explored for its potential application in radiology. This study assesses ChatGPT-4V's diagnostic performance in identifying various types of intracranial hemorrhages in non-contrast cranial computed tomography (CT) images.

## METHODS

Intracranial hemorrhages were presented to ChatGPT using the clearest 2D imaging slices. The first question, "Q1: Which imaging technique is used in this image?" was asked to determine the imaging modality. ChatGPT was then prompted with the second question, "Q2: What do you see in this image and what is the final diagnosis?" to assess whether the CT scan was normal or showed pathology. For CT scans containing hemorrhage that ChatGPT did not interpret correctly, a follow-up question—"Q3: There is bleeding in this image. Which type of bleeding do you see?"—was used to evaluate whether this guidance influenced its response.

## RESULTS

ChatGPT accurately identified the imaging technique (Q1) in all cases but demonstrated difficulty diagnosing epidural hematoma (EDH), subdural hematoma (SDH), and subarachnoid hemorrhage (SAH) when no clues were provided (Q2). When a hemorrhage clue was introduced (Q3), ChatGPT correctly identified EDH in 16.7% of cases, SDH in 60%, and SAH in 15.6%, and achieved 100% diagnostic accuracy for hemorrhagic cerebrovascular disease. Its sensitivity, specificity, and accuracy for Q2 were 23.6%, 92.5%, and 57.4%, respectively. These values improved substantially with the clue in Q3, with sensitivity rising to 50.9% and accuracy to 71.3%. ChatGPT also demonstrated higher diagnostic accuracy in larger hemorrhages in EDH and SDH images.

## CONCLUSION

Although the model performs well in recognizing imaging modalities, its diagnostic accuracy substantially improves when guided by additional contextual information.

## CLINICAL SIGNIFICANCE

These findings suggest that ChatGPT's diagnostic performance improves with guided prompts, highlighting its potential as a supportive tool in clinical radiology.

## KEYWORDS

Artificial intelligence, intracranial hemorrhages, ChatGPT, computed tomography, hematoma

Corresponding author: Olga Bayar-Kapıcı

E-mail: olgasahbayar@gmail.com

Received 24 May 2025; revision requested 10 June 2025; accepted 28 June 2025.



Epub: 21.07.2025

Publication date: 02.01.2026

DOI: 10.4274/dir.2025.253456

**A**rtificial intelligence (AI) is increasingly being used across various fields to assist humans in quickly accessing information and supporting decision-making processes.<sup>1</sup> One of the subcategories of AI, large language models (LLMs), is a type of generative AI capable of processing, understanding, and generating human knowledge. LLMs are trained using self-supervised learning, which enables them to predict missing or hidden elements within a text.<sup>2</sup> Among these LLMs, Chat Generative Pre-trained Transformer (ChatGPT) is built on the GPT-4 architecture. ChatGPT is a text-based model that supports decision-making

You may cite this article as: Bayar-Kapıcı O, Altunışık E, Musabeyoğlu F, Dev Ş, Kaya Ö. Artificial intelligence in radiology: diagnostic sensitivity of ChatGPT for detecting hemorrhages in cranial computed tomography scans. *Diagn Interv Radiol.* 2026;32(1):27-32.

across a wide range of domains. OpenAI's ChatGPT model operates using a supervised learning process in which it predicts the next element in a sequence of text.<sup>3</sup>

In November 2023, OpenAI updated the ChatGPT model to GPT-4V, which introduced the ability not only to communicate through text but also to interpret and generate images. GPT-4V is an enhanced version of ChatGPT with capabilities for processing visual data, allowing it to analyze and comment on images. This development expands its utility beyond text-based tasks, enabling performance in image-related contexts as well.<sup>4</sup> However, the visual capabilities of ChatGPT remain in development and currently have certain limitations.

There has been considerable discussion in the literature regarding the potential use of AI in radiology.<sup>5</sup> In acute settings, where rapid decision-making is critical, AI's ability to detect pathologies in radiological imaging may help reduce patient morbidity and mortality. It has been proposed that ChatGPT, in particular, could support the diagnosis of patients with time-sensitive conditions such as stroke.<sup>6</sup> In such cases, where timely intervention substantially lowers the risk of long-term disability, ChatGPT's diagnostic capabilities could offer valuable support.

In this study, we presented ChatGPT-4V with computed tomography images (CTI) of epidural hematoma (EDH), subdural hematoma (SDH), subarachnoid hemorrhage (SAH), hemorrhagic cerebrovascular disease (HSVD), and normal brain scans [non-contrast CT (NCT)], and asked it a series of questions. We aimed to evaluate its diagnostic

sensitivity and specificity in detecting hemorrhages based on the responses it provided.

## Methods

### Study design

This study was conducted at the Neurology Clinic of Gaziantep City Hospital. Approval was obtained from the Non-Interventional Clinical Research Ethics Committee of Gaziantep City Hospital (IRB number: 159/2025, decision date: March 19, 2025). All participants provided signed informed consent.

Participants were required to meet specific inclusion criteria. Only adults aged 18 years and older with available cranial CT scans were included. These scans needed to show either intracranial hemorrhages—such as EDH, SDH, SAH, or HSVD—or a normal brain. Additionally, the CTIs had to be clear and non-contrast, enabling accurate detection and classification of hemorrhages.

Exclusion criteria included pediatric patients under the age of 18, as well as participants with unclear CTIs or imaging artifacts that interfered with hemorrhage detection. Patients with a history of cognitive impairment or conditions that prevented them from providing informed consent were also excluded, as were individuals with a history of brain trauma or prior neurosurgical procedures, as these could influence the interpretation of current CT findings. Finally, cases involving chronic hemorrhages, where the acute nature of the pathology could not be confirmed, were excluded. A flowchart of the study is presented in Figure 1.

### Chat Generative Pre-trained Transformer assessment

Intracranial hemorrhages were presented to ChatGPT using the clearest 2D imaging slices. In Q1, ChatGPT was asked to identify the imaging technique (Figure 2). In Q2, it was prompted to determine whether the CTI was normal or to identify any pathology, if present (Figure 3). For scans showing hemorrhage that ChatGPT failed to interpret accurately, a follow-up prompt—Q3—was used to assess whether its response changed when guided to identify the type of bleeding (Figure 4).

### Statistical analysis

The analysis of ChatGPT's diagnostic performance was conducted using descriptive statistics to evaluate the model's success rates in answering questions related to

various hemorrhage types (EDH, SDH, SAH, HSVD, and NCT) presented in cranial CT scans. The success rates for each question (Q1, Q2, Q3) were calculated and reported as frequencies and percentages for each condition.

To determine the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of the model in identifying hemorrhages, 2 × 2 contingency tables were constructed for each question. Sensitivity (true positive rate) and specificity (true negative rate) were calculated for Q2 and Q3 to evaluate ChatGPT's diagnostic performance in detecting hemorrhages.

Further analysis involved comparing sensitivity and specificity across different hemorrhage types and questions (Q1, Q2, Q3). Additionally, the relationship between hemorrhage size and diagnostic accuracy was examined using *P* values derived from statistical tests (e.g., Mann–Whitney *U* test) to assess whether hemorrhage size influenced diagnostic outcomes.

## Results

ChatGPT correctly identified the imaging technique (Q1) in all images. When asked what it saw in the image without any clues (Q2), ChatGPT failed to correctly diagnose EDH, SDH, and SAH. However, when the image was identified as showing hemorrhage and the task was to determine the type, ChatGPT correctly identified EDH in 16.7%, SDH in 60%, and SAH in 15.6% of cases. For HSVD, ChatGPT achieved an 86.7% correct diagnosis rate in Q2 and reached 100% diagnostic accuracy in Q3. It also correctly identified negative findings in 92.5% of normal CT scans (Table 1 and Figure 5).

The sensitivity and specificity of ChatGPT for detecting intracranial hemorrhages are summarized in Table 2. For Q2, sensitivity was 23.6%, specificity was 92.5%, PPV was 76.5%, NPV was 53.8%, and overall accuracy was 57.4%. With the diagnostic clue provided in Q3, sensitivity increased to 50.9%, PPV to 87.5%, NPV to 64.5%, and accuracy to 71.3%.

The relationship between correct diagnoses of EDH and SDH in Q3 and hemorrhage size is shown in Table 3. According to these results, in EDH and SDH images, the hemorrhage size was statistically significantly larger in cases correctly diagnosed by ChatGPT compared with false negatives (*P* = 0.038 and *P* = 0.030, respectively).

## Discussion

The main findings of this study are as follows: (i) ChatGPT correctly identified the imaging modality in all images; (ii) it failed to provide accurate diagnoses in cases of EDH, SDH, and SAH, with the exception of HSVD; (iii) it was able to generate correct diagnoses when appropriately guided; and (iv) in EDH and SDH images, the hemorrhage diameter was larger in cases where ChatGPT provided the correct diagnosis.

This study evaluated the diagnostic capabilities of ChatGPT in identifying various types of intracranial hemorrhages using non-contrast cranial CTIs. The results highlight both the potential and the current

limitations of this large language and vision model in the context of neuroimaging interpretation.

The first key finding is that ChatGPT successfully identified the imaging modality as CT in 100% of cases. This suggests that the model is reliably capable of recognizing basic imaging types, even when presented with isolated slices and no clinical context. However, when tasked with identifying specific pathologies—particularly acute hemorrhages—its diagnostic performance was notably limited. The model was only able to correctly diagnose HSVD with high accuracy, whereas it consistently failed to detect EDH, SDH, and SAH without guidance.

These findings are important, as they reveal that while ChatGPT possesses a degree of image interpretation capacity, its baseline performance in detecting life-threatening hemorrhages remains suboptimal. A key secondary observation, however, is that the model's diagnostic accuracy improved considerably when guided with targeted questions (Q3). Prior research has suggested that LLMs such as ChatGPT tend to perform better in complex clinical tasks when questions are framed in open-ended or context-rich formats, which enhance the relevance and depth of their responses.<sup>7</sup>

Supporting our findings, a recent study by Kahalian et al.<sup>8</sup> evaluated ChatGPT-4V's diagnostic performance in interpreting oral and maxillofacial radiographic images. The authors reported that the correct pre-diagnosis rate was only 30.7% when no clues were provided, but this rate substantially increased to 56.9% with the inclusion of structured prompts, such as internal lesion features or anatomical context. These results confirm that providing domain-relevant cues can substantially enhance the diagnostic accuracy of GPT-4V in medical imaging tasks. Notably, similar to our study, the authors found that the model struggled to differentiate closely located anatomical structures and failed to generate comprehensive differential diagnoses in complex cases. This parallel reinforces the conclusion that, while ChatGPT-4V demonstrates baseline interpretive ability, its effective use in clinical radiology depends heavily on contextual scaffolding and targeted prompting strategies.

Recent literature has further emphasized the growing potential of LLMs in radiology, highlighting their capacity to support tasks ranging from protocol selection to diagnostic reasoning and structured reporting. Akinci D'Antonoli et al.<sup>9</sup> provided a comprehensive overview of how LLMs, such as GPT-4, may be integrated into radiological workflows to improve clinical decision-making.

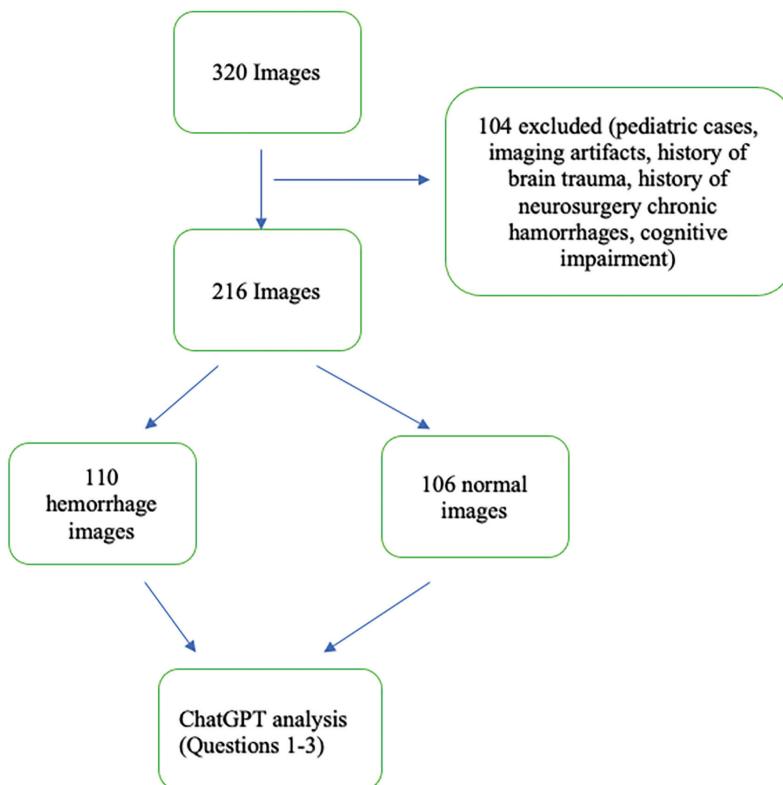
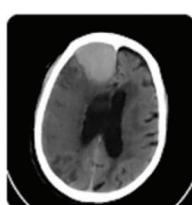


Figure 1. Flowchart of the study. ChatGPT, Chat Generative Pre-trained Transformer.



Which imaging technique is used in this image?

The imaging technique used in this image is **non-contrast cranial computed tomography (CT)**.

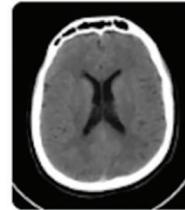
Figure 2. Q1: Which imaging technique is used in this image?

ing and enhance the efficiency of data interpretation. Although our study demonstrates that GPT-4V continues to underperform in detecting subtle hemorrhagic pathologies on cranial CT scans—particularly in the absence of contextual prompts—these broader applications suggest that LLMs may still contribute meaningfully when used for textual analysis, report structuring, or as conversational assistants in radiology departments. Future iterations of such models, especially those fine-tuned for radiological image data and integrated with clinical metadata, may

hold transformative potential in diagnostic radiology.

In contrast to our findings, which revealed diagnostic limitations of ChatGPT on cranial CTs, Kuzan et al.<sup>10</sup> observed improved performance in stroke diagnosis when diffusion-weighted imaging (DWI) magnetic resonance imaging (MRI) was utilized. In their study, ChatGPT-4V demonstrated a sensitivity of 79.5% and a specificity of 84.9% in detecting acute ischemic stroke using DWI and apparent diffusion coefficient maps. Although our results showed that ChatGPT-4V

struggled particularly with identifying EDH, SDH, and SAH, its success in HSVD cases and its improvement after guided prompts suggest that diagnostic performance is strongly influenced by the nature and clarity of radiological findings. The relatively high accuracy reported by Kuzan et al.<sup>10</sup> may be attributed to the more conspicuous radiologic features of diffusion restriction on MRI, compared with the often subtle or variable appearance of hemorrhages on CT. These findings underscore the importance of tailoring AI applications to specific imaging modalities



What do you see in this image? What is your final diagnosis?



This axial non-contrast CT scan of the brain appears relatively normal. Here's what I observe:

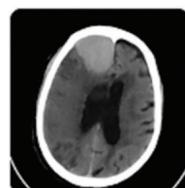
### Key findings:

- **Symmetric lateral ventricles** without midline shift
- **No hyperdense (acute blood)** or hypodense (edema or infarct) areas
- **No mass effect**, sulcal effacement, or herniation signs
- **Gray-white matter differentiation** is preserved
- **No evidence of skull fracture or extracranial soft tissue swelling**

### Final Diagnosis:

Normal cranial CT — no acute intracranial pathology is visible in this slice.

Figure 3. Q2: What do you see in this image, and what is the final diagnosis? CT, computed tomography.



There is a bleeding in this image. Which type of bleeding do you see?

This CT image shows a **hyperdense (bright) crescent-shaped area** along the **left cerebral hemisphere**, causing a midline shift and compression of the lateral ventricle. These findings are characteristic of a **subdural hematoma (SDH)**.

Figure 4. Q3: There is bleeding in this image. Which type of bleeding do you see?

and reinforce the potential of ChatGPT as a supportive tool when used within defined clinical and technical contexts.

Furthermore, the study showed that in EDH and SDH cases, the hemorrhage diameter was substantially greater in the true pos-

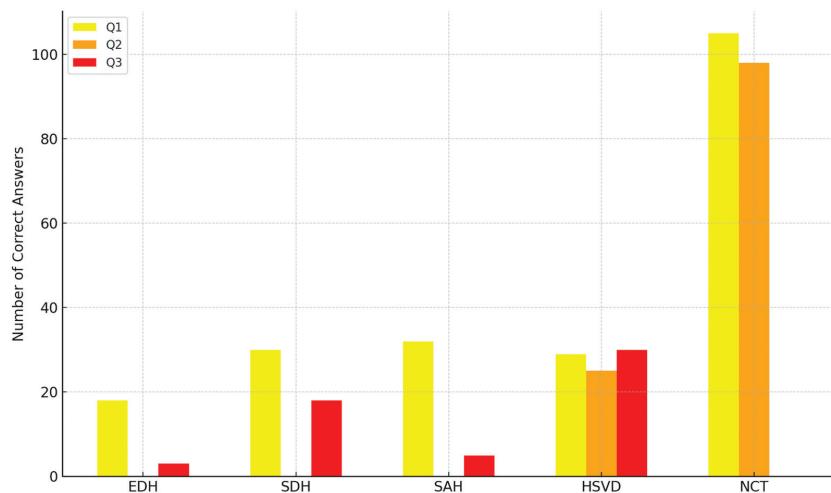
itive group than in the false negative group. This suggests that the model may be more adept at recognizing larger and more prominent pathologies and may struggle with subtle or borderline findings. This size-related variability in diagnostic accuracy has impor-

tant implications for clinical practice, where early detection of small-volume hemorrhages is often critical for timely intervention.

A recent study by Koyun et al.<sup>11</sup> evaluated the diagnostic capabilities of ChatGPT-4V in identifying various types of intracranial hemorrhages on NCT and reported promising results, with a sensitivity of 79.2% and an accuracy of 68.3% in hemorrhage detection. However, their findings also revealed notable limitations in localizing hemorrhages and identifying subarachnoid and epidural types, particularly in the absence of clear density differences. These results are consistent with our study, which demonstrated that ChatGPT's performance was markedly better for hemorrhages with larger diameters and distinct features (e.g., HSVD), whereas its diagnostic accuracy was considerably lower in cases of EDH, SDH, and SAH. Notably, both studies found that the model was highly consistent in identifying the imaging modality but often failed in complex classification tasks without tailored prompting. Together, these findings highlight the current strengths and limitations of general-purpose LLMs in radiologic interpretation and reinforce the need for multimodal training and task-specific tuning for clinical use.

One of the most comprehensive assessments of GPT-4V's performance in neuro-imaging was recently conducted by Zhang et al.<sup>12</sup>, who analyzed the model's ability to detect and annotate cerebral hemorrhages on non-contrast cranial CTs. In their retrospective evaluation of 208 CT scans, GPT-4V achieved an overall identification completeness of 72.6%, with the highest performance observed in epidural and intraparenchymal hemorrhages (89.0% and 86.9%, respectively). However, it showed substantially lower performance in chronic SDH and SAH, mirroring the diagnostic gaps also noted in our study. Their results also indicated that GPT-4V was more accurate in identifying massive hemorrhages than minor ones, supporting our finding that larger bleeding volumes in EDH and SDH were associated with better diagnostic accuracy. Together, these findings underscore the model's dependence on the visual salience of hemorrhagic lesions and reaffirm the need for multimodal refinement and clinical oversight if GPT-4V is to be integrated into routine radiologic workflows.

In addition to our findings on image-based diagnostic limitations, recent research has also highlighted concerns regarding the textual output of AI models. Güll et al.<sup>13</sup> conducted a cross-sectional study eval-



**Figure 5.** Number of correct responses by ChatGPT. ChatGPT, Chat Generative Pre-trained Transformer; EDH, epidural hematoma; SDH, subdural hematoma; SAH, subarachnoid hemorrhage; HSVD, hemorrhagic cerebrovascular disease; NCT, non-contrast computed tomography.

**Table 1.** ChatGPT's success rates in answering the questions

|               | Q1<br>n (%) | Q2<br>n (%) | Q3<br>n (%) |
|---------------|-------------|-------------|-------------|
| EDH (n = 18)  | 18 (100)    | 0 (0)       | 3 (16.7)    |
| SDH (n = 30)  | 30 (100)    | 0 (0)       | 18 (60)     |
| SAH (n = 32)  | 32 (100)    | 0 (0)       | 5 (15.6)    |
| HSVD (n = 30) | 30 (100)    | 26 (86.7)   | 30 (100)    |
| NCT (n = 106) | 106 (100)   | 98 (92.5)   |             |

ChatGPT, Chat Generative Pre-trained Transformer; EDH, epidural hematoma; SDH, subdural hematoma; SAH, subarachnoid hemorrhage; HSVD, hemorrhagic cerebrovascular disease; NCT, normal cranial computed tomography images.

**Table 2.** Sensitivity and specificity analysis of ChatGPT

| Q2 | TP | TN | FP | FN | Sensitivity | Specificity | PPV   | NPV    | Accuracy |
|----|----|----|----|----|-------------|-------------|-------|--------|----------|
| n  | 26 | 98 | 8  | 84 | 26/110      | 98/106      | 26/34 | 98/182 | 124/216  |
| %  |    |    |    |    | 23.6        | 92.5        | 76.5  | 53.8   | 57.4     |
| Q3 | TP | TN | FP | FN | Sensitivity | Specificity | PPV   | NPV    | Accuracy |
| n  | 56 | 98 | 8  | 54 | 56/110      | 98/106      | 56/64 | 98/152 | 154/216  |
| %  |    |    |    |    | 50.9        | 92.5        | 87.5  | 64.5   | 71.3     |

Sensitivity = TP / (TP + FN), Specificity = TN / (TN + FP), PPV = TP / (TP + FP), NPV = TN / (TN + FN), Accuracy = (TP + TN) / (TP + TN + FP + FN).

TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; NPV, negative predictive value.

**Table 3.** Comparison of ChatGPT's results based on hemorrhage size

| Q3                | FN median (min.–max.) | TP median (min.–max.) | P value |
|-------------------|-----------------------|-----------------------|---------|
| EDH diameter (mm) | 10 (3–32)             | 17 (16–20)            | 0.038   |
| SDH diameter (mm) | 19 (10–35)            | 22.5 (12–41)          | 0.030   |

The Mann–Whitney U test was used. A P value <0.05 was considered statistically significant.

EDH, epidural hematoma; SDH, subdural hematoma; mm, millimeter; FN, false negatives; TP, true positives; min.–max., minimum–maximum.

uating the quality, reliability, and readability of ChatGPT, Bard, and Perplexity responses to patient-centered questions on SDH. They found that all three AI tools produced answers that substantially exceeded the recommended sixth-grade reading level, making the content difficult for general users to understand. Moreover, ChatGPT's responses were rated lower in readability than Bard and Perplexity, and its DISCERN and JAMA quality scores indicated deficiencies in transparency, citation, and clarity. These findings reinforce the need for optimization not only in visual diagnostic performance, as shown in our study, but also in natural language output, especially for patient education. They further support the idea that while LLMs show potential in clinical contexts, their use must be accompanied by careful oversight and task-specific calibration to avoid misleading or inaccessible information.

These limitations align with the understanding that GPT-4V, although capable of processing visual inputs, has not been trained on annotated radiologic datasets and lacks the spatial learning capabilities typical of convolutional neural networks.<sup>14</sup> As a result, its ability to detect subtle radiographic features remains inherently limited.

Nevertheless, the model's capacity to engage in guided reasoning and improve diagnostic performance when provided with contextual prompts presents promising potential. With further domain-specific training and integration of multimodal clinical data—such as patient history, symptoms, and laboratory results—LLMs may evolve into useful adjunct tools in emergency and diagnostic radiology.

This study has several limitations that should be considered. First, the sample size for each hemorrhage type was relatively small, which may limit the generalizability of the results. Second, the accuracy of ChatGPT-4V was influenced by the clarity and quality of the CTIs, as some scans contained artifacts or poor resolution, potentially affecting performance. Third, the retrospective nature of the study means the findings are based on historical data, and real-time clinical valida-

tion is needed to confirm the model's practical utility. Additionally, although ChatGPT-4V showed improved diagnostic accuracy when given clues, its performance in more complex or subtle hemorrhage cases remains uncertain, suggesting the need for further refinement. Lastly, the lack of comparison with other AI models or radiologists limits the ability to fully assess ChatGPT's relative effectiveness in diagnosing intracranial hemorrhages.

In conclusion, while ChatGPT demonstrates basic competence in identifying imaging modalities and limited ability in hemorrhage detection—particularly in HSVD—it is not yet suitable for autonomous radiologic interpretation. However, its interactive design and improved performance under guidance suggest that LLMs may serve a valuable supportive role in the future, particularly when embedded within supervised or hybrid diagnostic systems.

#### Footnotes

#### Conflict of interest disclosure

The authors declared no conflicts of interest.

#### References

1. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Assessing artificial intelligence for clinical decision-making. *Front Digit Health.* 2021;3:645232. [\[Crossref\]](#)
2. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *AJR Am J Roentgenol.* 2023;221(3):373-376. [\[Crossref\]](#)
3. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J Pediatr Urol.* 2023;19(5):598-604. [\[Crossref\]](#)
4. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol.* 2025;35(1):506-516. [\[Crossref\]](#)
5. Boeken T, Feydy J, Lecler A, et al. Artificial intelligence in diagnostic and interventional radiology: where are we now? *Diagn Interv Imaging.* 2023;104(1):1-5. [\[Crossref\]](#)
6. Gilotra K, Swarna S, Mani R, Basem J, Dashti R. Role of artificial intelligence and machine learning in the diagnosis of cerebrovascular disease. *Front Hum Neurosci.* 2023;17:1254417. [\[Crossref\]](#)
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198. [\[Crossref\]](#)
8. Kahalian S, Rajabzadeh M, Özbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Med (Plovdiv).* 2024;66(6):863-868. [\[Crossref\]](#)
9. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol.* 2024;30(2):80-90. [\[Crossref\]](#)
10. 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(3):187-195. [\[Crossref\]](#)
11. Koyun M, Cevval ZK, Reis B, Ece B. Detection of intracranial hemorrhage from computed tomography images: diagnostic role and efficacy of ChatGPT-4o. *Diagnostics (Basel).* 2025;15(2):143. [\[Crossref\]](#)
12. Zhang D, Ma Z, Gong R, et al. Using natural language processing (GPT-4) for computed tomography image analysis of cerebral hemorrhages in radiology: retrospective analysis. *J Med Internet Res.* 2024;26:e58741. [\[Crossref\]](#)
13. Güll Ş, Erdemir İ, Hancı V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT, BARD, and perplexity responses. *Medicine (Baltimore).* 2024;103(18):e38009. [\[Crossref\]](#)
14. OpenAI. GPT-4 technical report. 2023. [\[Crossref\]](#)