Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242892

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright© Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke

Beyza Nur Kuzan[1]
İsmail Meşe[2]
Servan Yaşar[3]
Taha Yusuf Kuzan[3]

[1]Kartal Dr. Lütfi Kırdar City Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]Üsküdar State Hospital, Clinic of Radiology, İstanbul, Türkiye

[3]Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

Stroke is a neurological emergency requiring rapid, accurate diagnosis to prevent severe consequences. Early diagnosis is crucial for reducing morbidity and mortality. Artificial intelligence (AI) diagnosis support tools, such as Chat Generative Pre-trained Transformer (ChatGPT), offer rapid diagnostic advantages. This study assesses ChatGPT's accuracy in interpreting diffusion-weighted imaging (DWI) for acute stroke diagnosis.

METHODS

A retrospective analysis was conducted to identify the presence of stroke using DWI and apparent diffusion coefficient (ADC) map images. Patients aged >18 years who exhibited diffusion restriction and had a clinically explainable condition were included in the study. Patients with artifacts that affected image homogeneity, accuracy, and clarity, as well as those who had undergone previous surgery or had a history of stroke, were excluded from the study. ChatGPT was asked four consecutive questions regarding the identification of the magnetic resonance imaging (MRI) sequence, the demonstration of diffusion restriction on the ADC map after sequence recognition, and the identification of hemispheres and specific lobes. Each question was repeated 10 times to ensure consistency. Senior radiologists subsequently verified the accuracy of ChatGPT's responses, classifying them as either correct or incorrect. We assumed a response to be incorrect if it was partially correct or suggested multiple answers. These responses were systematically recorded. We also recorded non-responses from ChatGPT-4V when it failed to provide an answer to a query. We assessed ChatGPT-4V's performance by calculating the number and percentage of correct responses, incorrect responses, and non-responses across all images and questions, a metric known as "accuracy." ChatGPT-4V was considered successful if it answered ≥80% of the examples correctly.

RESULTS

A total of 530 diffusion MRI, of which 266 were stroke images and 264 were normal, were evaluated in the study. For the initial query identifying MRI sequence type, ChatGPT-4V's accuracy was 88.3% for stroke and 90.1% for normal images. For detecting diffusion restriction, ChatGPT-4V had an accuracy of 79.5% for stroke images, with a 15% false positive rate for normal images. Regarding identifying the brain or cerebellar hemisphere involved, ChatGPT-4V correctly identified the hemisphere in 26.2% of stroke images. For identifying the specific brain lobe or cerebellar area affected, ChatGPT-4V had a 20.4% accuracy for stroke images. The diagnostic sensitivity of ChatGPT-4V in acute stroke was found to be 79.57%, with a specificity of 84.87%, a positive predictive value of 83.86%, a negative predictive value of 80.80%, and a diagnostic odds ratio of 21.86.

CONCLUSION

Despite limitations, ChatGPT shows potential as a supportive tool for healthcare professionals in interpreting diffusion examinations in stroke cases, where timely diagnosis is critical.

CLINICAL SIGNIFICANCE

ChatGPT can play an important role in various aspects of stroke cases, such as risk assessment, early diagnosis, and treatment planning.

KEYWORDS

Ischemic stroke, ChatGPT, magnetic resonance imaging, diffusion-weighted imaging

Corresponding author: Beyza Nur Kuzan

E-mail: drbeyzauzun@hotmail.com

Received 06 June 2024; revision requested 02 July 2024; accepted 28 July 2024.

Epub: 02.09.2024

Publication date: 28.04.2025

DOI: 10.4274/dir.2024.242892

You may cite this article as: 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.

Artificial intelligence (AI) is a set of applications that can be used in almost any field to support human power and decision-making processes. Within AI, there are several subcategories, including deep learning, machine learning (ML), and generative AI, with the latter gaining significant popularity recently. The historical progression of AI, specifically generative and multimodal AI, can be traced back to the early 20th century with the development of the Markov chain model in 1906, which laid the foundation for probabilistic methods in AI.[1] Significant advancements occurred in the mid-20th century with the rise of natural language processing and ML, leading to early chatbots, such as ELIZA, in the 1960s.[2] Notable milestones include the Turing test in 1950, which set a benchmark for machine intelligence, and the creation of rule-based chatbots in the 1960s and 1970s.[1] The integration of deep learning in the early 2000s led to the development of large language models (LLMs), such as OpenAI's Chat Generative Pre-trained Transformer (ChatGPT) and Google's Bard, which utilize transformer neural network architectures. These sophisticated conversational agents are advanced AI systems trained on extensive datasets. These models predict the next word in a sentence, enabling them to generate coherent and contextually relevant text based on the input they receive.[1]

The field of radiology is undergoing a significant transformation with the introduction of AI. This transformation includes AI-powered tools and plug-ins that can analyze large multi-view datasets, identifying patterns that are not easily detected by the human eye. AI algorithms can also assist radiologists by automating routine tasks.[3,4] These innovations have led to improved image quality, reductions in scan times, and the development of predictive analytics for patient outcomes. Another critical aspect of this AI-driven transformation is the ability to personalize patient care.[5]

Main points

- Chat Generative Pre-trained Transformer (ChatGPT) is a tool that can potentially assist healthcare professionals in diagnosing diseases.
- Although ChatGPT offers rapid and comprehensive responses, as well as convenient accessibility, it also has certain drawbacks, including sometimes inconsistent outputs and the necessity for supervision.
- The findings of this study indicate that, despite its current limitations, ChatGPT demonstrated a 79.5% success rate in determining diffusion restriction in stroke cases.

The role of AI becomes even more critical in situations where the timing of diagnosis affects morbidity and mortality, such as stroke cases. Rapid imaging is crucial in stroke cases because timely intervention can significantly reduce the risk of long-term disability and improve patient outcomes.[6] AI creates a diagnostic advantage in these emergency cases due to its easy accessibility and rapid decision-making features.[7] It offers a promising solution to bridge the gap, particularly in cases where the limited availability of radiologists presents a significant challenge.[8]

In November 2023, OpenAI unveiled a groundbreaking update to ChatGPT with the introduction of its Generative Pre-trained Transformer 4, enhanced with vision capabilities, known as GPT-4V.[9,10] This update transforms ChatGPT from merely a tool for textual analysis into a versatile assistant capable of handling a wide range of tasks that require an understanding of both language and visual data. In a recent article by Kim et al.[11], the authors used ChatGPT-4V to interpret radiology examinations, despite it scoring lower than the students. In another article by Deng et al.[12], it was found to have limited accuracy and precision, inconsistent performance, and a tendency to "hallucinate". Despite these reports, the use of ChatGPT-4V in radiology, especially in stroke imaging, remains largely unexplored. Because of its rapid interpretation and practical accessibility, the use of ChatGPT in the diagnosis of stroke should be investigated in large case series. Clinical application and dissemination of ChatGPT by verifying its diagnostic performance and suitability for stroke diagnosis will develop this field.

In this study, we aim to evaluate the diagnostic accuracy and effectiveness of interpreting diffusion-weighted imaging (DWI) using ChatGPT in the diagnosis of acute stroke. Our method involves a structured approach to posing specific questions of varying difficulty, each designed to address different aspects of image interpretation in stroke imaging, from identifying the magnetic resonance imaging (MRI) sequence to pinpointing the specific location and lobe of the acute infarct.

Methods

Patient selection

This study was conducted in accordance with ethical standards and was approved by the Institutional Review Board of Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital (approval number: 33/14.02.2024). The requirement for informed written consent was waived due to the retrospective nature of the study.

A retrospective analysis was conducted on DWI and apparent diffusion coefficient (ADC) maps acquired between January 2022 and January 2024 using the institutional Picture Archiving and Communication System (Simplex PACS, Ankara, Türkiye). The patients presenting with acute stroke symptoms (weakness in the half of the body, difficulty in understanding and speaking, facial asymmetry, diplopia, and vision loss) were evaluated in the emergency unit, and those in which stroke was considered in the preliminary diagnosis were imaged with diffusion MRI. In patients whose symptoms regressed during 24-hour observation, the diagnosis of transient ischemic attack (TIA) was considered, and these patients were not included in the study. The inclusion criteria were adults aged >18 years who had diffusion restriction and explained the clinical condition. Exclusion criteria were the presence of image artifacts that could affect the interpretation of the scans, previous history of stroke or neurosurgical intervention, pediatric patients aged <18 years, or lacunar infarcts <1 cm; patients diagnosed with TIA were excluded for the clarity, reliability, and homogeneity of the analyzed data. Images of patients without diffusion restriction and stroke symptoms in diffusion-weighted examinations were included in the study as normal images. In the study, 530 images, 266 stroke images, and 264 normal images were evaluated.

Radiologist assessment

All images were obtained using two identical 1.5T MRI (GE Healthcare SIGNA™) devices of the same model. DWI and ADC map images of patients meeting the inclusion criteria were evaluated independently by two radiologists with 8 and 9 years of experience in this field. The assessments were conducted by consensus, with both radiologists collaborating to determine the presence or absence of diffusion restriction. This consensus-based approach was performed to provide a reliable reference for ChatGPT evaluations. The imaging parameters were standardized across all scans according to the MRI protocol, including a *b* value of 0–1,000 s/mm^2, TR/TE of 5,000/60 ms, a slice thickness of 5 mm, and a matrix size of 128 × 128. A total of 530 images were included in the study, comprising 264 images from patients with normal DWI and ADC findings and 266 images from patients diagnosed with acute stroke based on DWI scans, exhibiting diffusion restriction in the DWI and ADC sequences (Figure 1).

ChatGPT-4V assessment of diffusion-weighted imaging scans

The selection criteria for MRI slices focused on those exhibiting the most representative areas of diffusion restriction. Random slices were selected if no diffusion restriction was present, prioritizing those with the highest probability of infarction, particularly in the middle cerebral artery region. High-quality images were chosen to ensure clarity in interpretation. The images used for input were in JPEG format, with a file size of approximately 500 kB each and a resolution of 512 × 512 pixels.

ChatGPT-4V was utilized to interpret the DWI scans. ChatGPT-4V can be influenced by file names or any hinted answers placed as text in the image, as it seems to draw context from them when generating responses.[12] Therefore, before starting, all information was deleted from the text and the image names were standardized, starting sequentially (Figure 2). DWI images were anonymized before being uploaded to the ChatGPT platform for interpretation using standardized prompts. The questions were asked for each scan, and prompts were in English, a language in which the language model demonstrated high comprehension capacity.[13] The four specific questions posed to ChatGPT-4V were carefully chosen to evaluate its ability to interpret DWI scans accurately (Figure 3). First, ChatGPT-4V was asked to identify the type of MRI sequence to ensure it correctly understood the image's context. Once the sequence type was identified, an additional ADC map was provided to check for diffusion restrictions. The last two questions tested ChatGPT-4V's ability to discern detailed anatomical structures and spatial orientation within the brain, which are crucial for precise medical interpretation.

Each question was asked 10 times for every image to ensure consistency in responses. Although a larger number of iterations could provide more comprehensive insights, it was impractical within the scope of the study. The session was restarted after each set of questions to ensure that ChatGPT was not influenced by prior interactions. The accuracy of ChatGPT's responses was subsequently verified by senior radiologists in a binary manner: either correct or incorrect. If a response was partially correct or suggested multiple answers, it was assumed to be incorrect. These responses were systematically recorded. If ChatGPT-4V did not provide an answer to a query, this non-response was also recorded. The performance of ChatGPT-4V was assessed using the number and percentage of correct responses, incorrect responses, and non-responses across all images and questions, referred to as "accuracy" (Figures 4-7). Due to consistency concerns, ChatGPT 4V was deemed successful only if it answered ≥80% of the examples correctly and was allowed to move on to the next question. If this threshold was not met, subsequent questions would not be asked, ensuring that only complete and accurate analyses were recorded. The success rate was calculated by dividing the number of correct answers by the total number of answers given.

Statistical analysis

The primary outcome measure was the accuracy and success rate of ChatGPT's re-

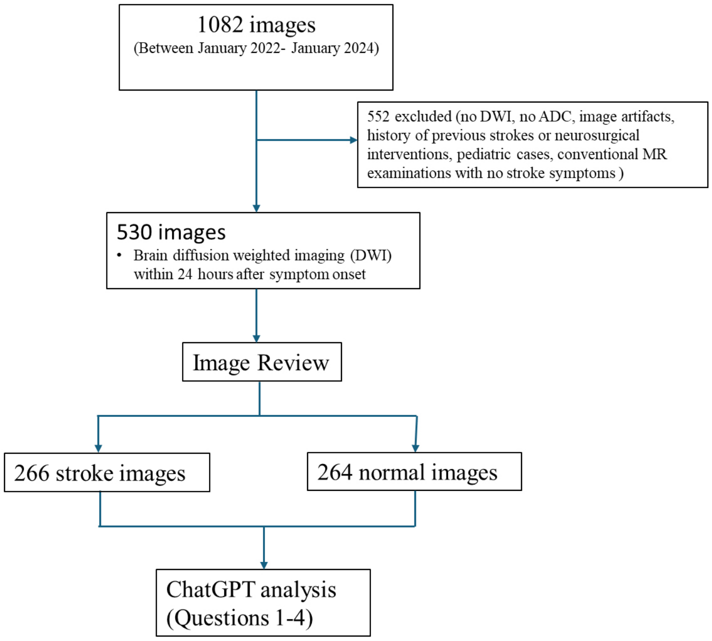
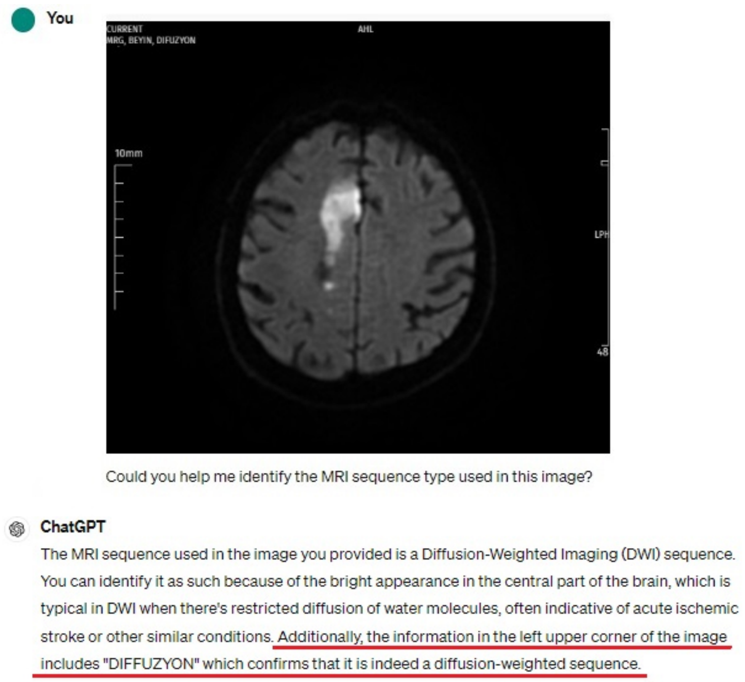

1082 images
(Between January 2022- January 2024)
552 excluded (no DWI, no ADC, image artifacts, history of previous strokes or neurosurgical interventions, pediatric cases, conventional MR examinations with no stroke symptoms)
530 images
• Brain diffusion weighted imaging (DWI) within 24 hours after symptom onset
Image Review
266 stroke images
264 normal images
ChatGPT analysis (Questions 1-4)

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



You

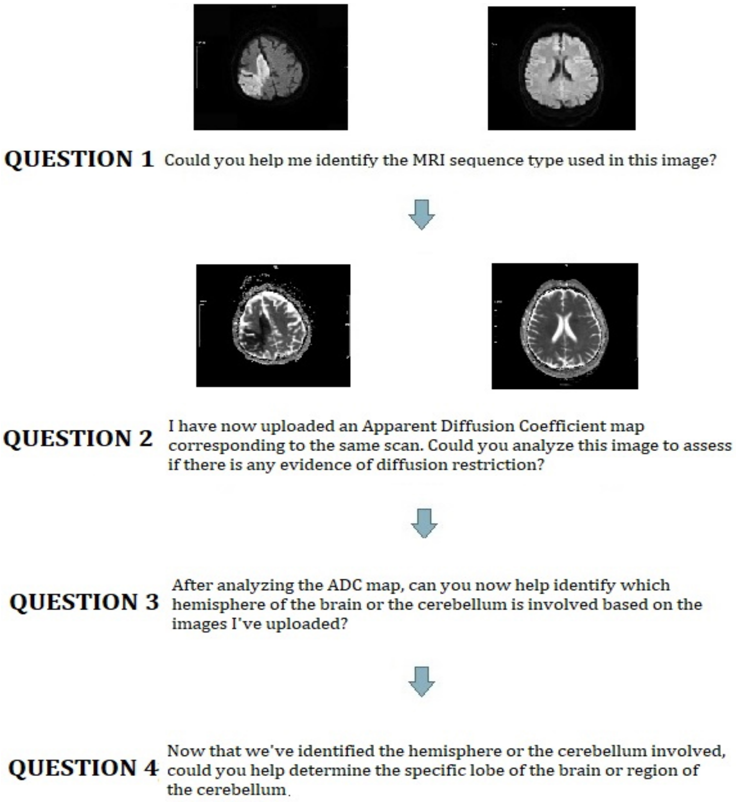
Could you help me identify the MRI sequence type used in this image?

ChatGPT

The MRI sequence used in the image you provided is a Diffusion-Weighted Imaging (DWI) sequence. You can identify it as such because of the bright appearance in the central part of the brain, which is typical in DWI when there's restricted diffusion of water molecules, often indicative of acute ischemic stroke or other similar conditions. Additionally, the information in the left upper corner of the image includes "DIFFUZYON" which confirms that it is indeed a diffusion-weighted sequence.

Figure 2. ChatGPT-4V can be influenced by filenames and image text as clues in interpretation. ChatGPT, Chat Generative Pre-trained Transformer.

sponses. The true positive (N_{TP}) and true negative (N_{TN}) are the number of patients correctly diagnosed as acute stroke and normal, respectively. In addition, normal cases wrongly diagnosed as stroke and incorrectly diagnosed stroke cases are assigned as (N_{FP}) and (N_{FN}), respectively. The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of ChatGPT in the diagnosis of acute stroke were calculated. The SPSS 23.0 (IBM Inc., Armonk, NY, USA) software package was used for statistical analysis.



QUESTION 1 Could you help me identify the MRI sequence type used in this image?

QUESTION 2 I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

QUESTION 3 After analyzing the ADC map, can you now help identify which hemisphere of the brain or the cerebellum is involved based on the images I've uploaded?

QUESTION 4 Now that we've identified the hemisphere or the cerebellum involved, could you help determine the specific lobe of the brain or region of the cerebellum.

Figure 3. Diagram of question texts and diffusion images asked to ChatGPT. ChatGPT, Chat Generative Pre-trained Transformer.

Results

In this retrospective study, the performance of ChatGPT-4V in interpreting DWI scans and ADC maps for a total of 530 images, including 266 stroke images and 264 normal images, was evaluated with various parameters. The results are divided into responses to four specific questions aimed at analyzing the capability of ChatGPT-4V in identifying critical aspects of DWI scans. Correct interpretations, incorrect interpretations, no responses, and success rates are shown in Table 1.

For the first question regarding the identification of the MRI sequence type, ChatGPT-4V accurately identified the MRI sequence in 235 images, resulting in an 88.3% success rate in the group of 266 stroke images. Overall, out of 2,660 interpretations, 2,098

Table 1. Performance analysis of ChatGPT-4V in interpreting diffusion-weighted imaging scans and apparent diffusion coefficient maps for stroke diagnosis

Group	Correct interpretations	Incorrect interpretations	No response	Success rate*
Question 1: Identification of the magnetic resonance imaging sequence type				
Stroke images (n = 266)	78.9% (2,098/2,660)	11.4% (305/2,660)	9.7% (257/2,660)	88.3%
Normal images (n = 264)	81.4% (2,148/2,640)	9.2% (244/2,640)	9.4% (248/2,640)	90.1%
Total images (n = 530)	79.7% (4,246/5,300)	10.4% (549/5,300)	9.5% (505/5,300)	89.2%
Question 2: Identification of diffusion restriction				
Stroke images (n = 235)	68.3% (1,605/2,350)	14.6% (343/2,350)	17.1% (402/2,350)	79.6%
Normal images (n = 238)	80.2% (1,909/2,380)	13.5% (320/2,380)	6.3% (151/2,380)	84.9%
Total images (n = 473)	74.4% (3,514/4,730)	14.0% (663/4,730)	11.7% (553/4,730)	82.2%
Question 3: The hemisphere of the brain or cerebellum involved				
Stroke images (n = 187)	32.4% (605/1,870)	53.5% (1,002/1,870)	14.1% (263/1,870)	26.2%
Question 4: The specific lobe of the brain or region of the cerebellum affected				
Stroke images (n = 49)	35.1% (172/490)	52.0% (255/490)	12.9% (63/490)	20.4%

*, achieved by attaining ≥80% correct interpretations in responded queries for each case; ChatGPT, Chat Generative Pre-trained Transformer.

were correct (78.9%), 257 did not receive a response (9.7%), and 305 were incorrect (11.4%). Similarly, in the group of 264 images with normal DWI findings, ChatGPT-4V successfully identified the sequence in 238 images (90.1%). Out of 2,640 interpretations for this group, 2,148 were correct (81.4%), 248 received no response (9.4%), and 244 were incorrect (9.2%).

In the second question concerning the identification of diffusion restriction, ChatGPT-4V successfully identified diffusion restriction in 187 out of 235 stroke images, indicating a 79.5% success rate for this subgroup. Out of 2,350 interpretations, 1,605 were correct (68.3%), 402 received no response (17.1%), and 343 were incorrect (14.6%). Conversely, ChatGPT-4V incorrectly identified diffusion restriction in 36 out of 238 normal images (15.1%), with 1,909 correct interpretations (indicating no diffusion restriction, 80.2%), 151 non-responses (6.3%), and 320 incorrect interpretations (13.5%).

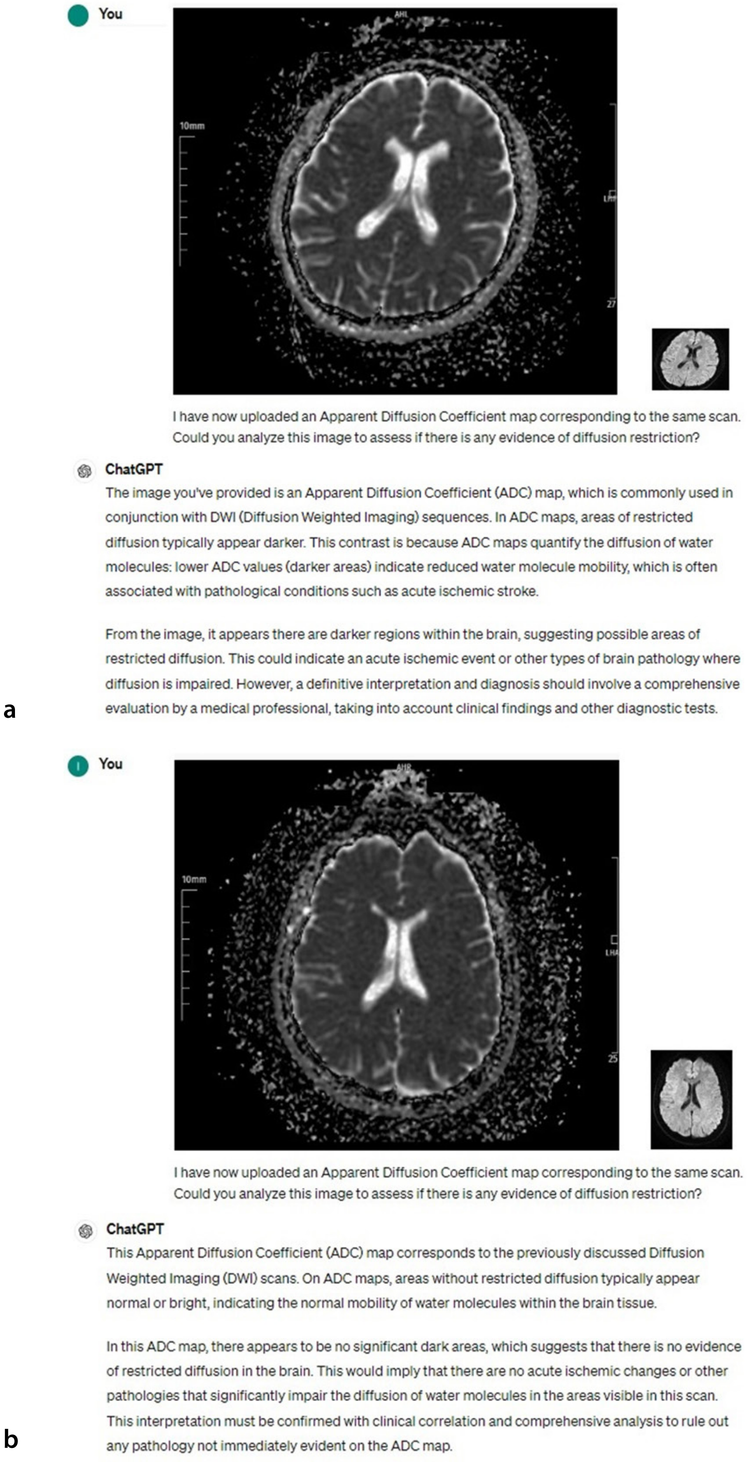
For the third question, regarding the hemisphere of the brain or cerebellum involved, ChatGPT-4V correctly identified the involved hemisphere in 49 out of 187 stroke images (26.2%). Out of 1,870 interpretations, 605 were correct (32.4%), 263 did not receive a response (14.1%), and 1,002 were incorrect (53.5%).

In the final question about the specific lobe of the brain or region of the cerebellum affected, ChatGPT-4V accurately identified the affected region in 10 out of 49 stroke images (20.4%). Out of 490 interpretations, 172 were correct (35.1%), 63 received no response (12.9%), and 255 were incorrect (52.0%). Further analysis revealed that ChatGPT-4V's interpretations were most successful for the frontal lobe (33.3%, 3 out of 9) and parietal lobe (30.0%, 3 out of 10), whereas its success rates for the temporal and occipital lobes were lower, at 15.0% (3 out of 20) and 10.0% (1 out of 10), respectively.

The diagnostic performance results obtained by comparing the images of stroke and normal with ChatGPT are shown in Table 2. Accordingly, 187 true positive interpretations and 48 false negative interpretations were made on the diffusion images of 235 stroke images. A total of 202 true negative interpretations and 38 false positive interpretations were made on the diffusion images of 238 normal images. Accordingly, the diagnostic sensitivity of ChatGPT was calculated as 79.57%, specificity as 84.87%, positive predictive value as 83.86%, negative predictive value as 80.80%, and diagnostic odds ratio as 21.86.

Discussion

The pivotal aspect of this study is the evaluation of ChatGPT-4V's ability to interpret DWI scans and ADC maps for stroke diagnosis. Our investigation reveals that AI, specifically advanced language models with enhanced vision capabilities, can contribute to the analysis of medical imaging in stroke imaging. The detailed analysis showed ChatGPT-4V's success in identifying MRI sequence types and assessing the presence of diffusion restriction, illustrating its utility in basic diagnostic tasks.



You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

The image you've provided is an Apparent Diffusion Coefficient (ADC) map, which is commonly used in conjunction with DWI (Diffusion Weighted Imaging) sequences. In ADC maps, areas of restricted diffusion typically appear darker. This contrast is because ADC maps quantify the diffusion of water molecules: lower ADC values (darker areas) indicate reduced water molecule mobility, which is often associated with pathological conditions such as acute ischemic stroke.

From the image, it appears there are darker regions within the brain, suggesting possible areas of restricted diffusion. This could indicate an acute ischemic event or other types of brain pathology where diffusion is impaired. However, a definitive interpretation and diagnosis should involve a comprehensive evaluation by a medical professional, taking into account clinical findings and other diagnostic tests.

You

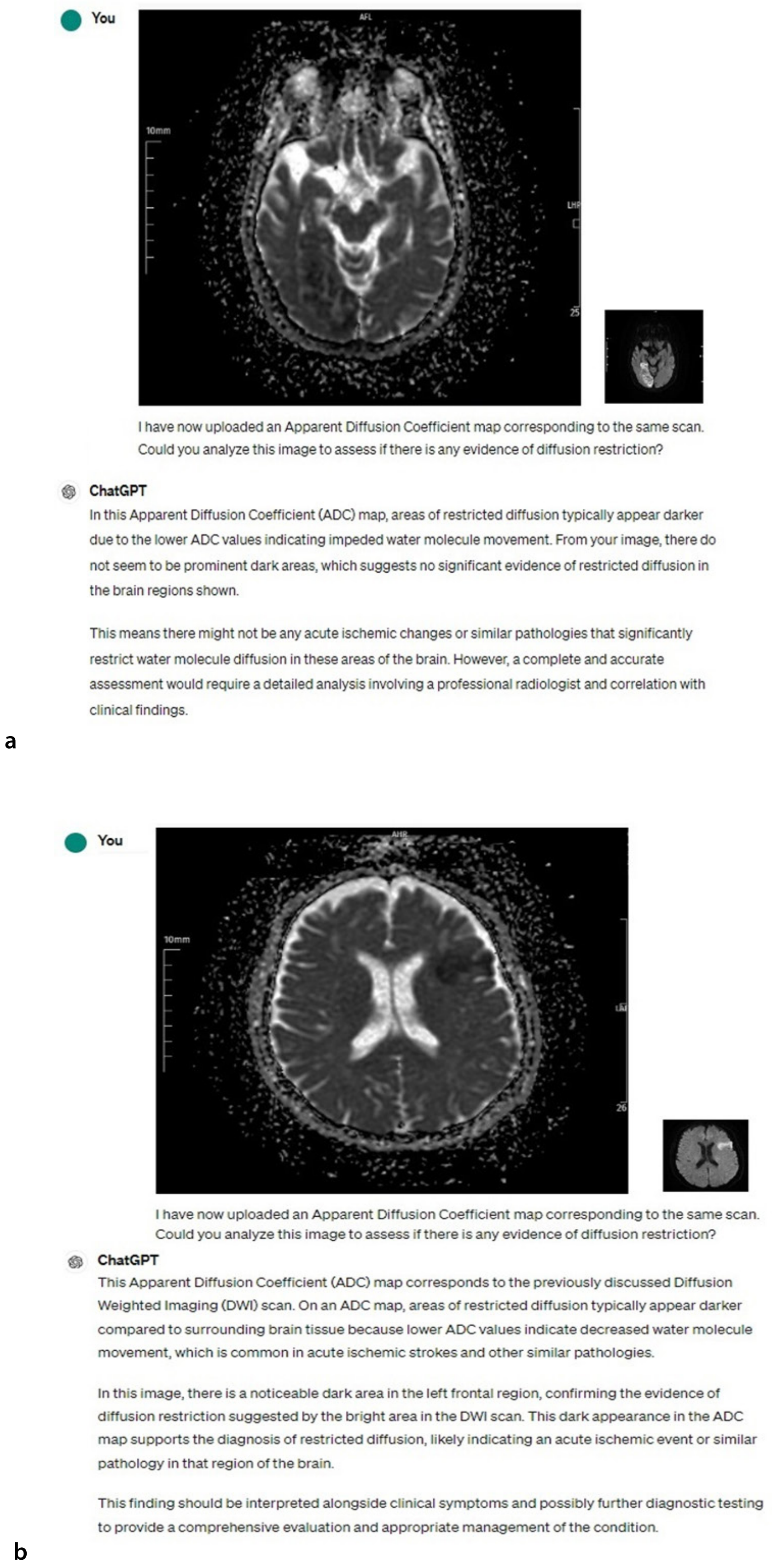
I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

This Apparent Diffusion Coefficient (ADC) map corresponds to the previously discussed Diffusion Weighted Imaging (DWI) scans. On ADC maps, areas without restricted diffusion typically appear normal or bright, indicating the normal mobility of water molecules within the brain tissue.

In this ADC map, there appears to be no significant dark areas, which suggests that there is no evidence of restricted diffusion in the brain. This would imply that there are no acute ischemic changes or other pathologies that significantly impair the diffusion of water molecules in the areas visible in this scan. This interpretation must be confirmed with clinical correlation and comprehensive analysis to rule out any pathology not immediately evident on the ADC map.

Figure 4. ChatGPT's interpretations **(a, b)** in apparent diffusion coefficient images without diffusion restriction for acute stroke. ChatGPT, Chat Generative Pre-trained Transformer.



You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

In this Apparent Diffusion Coefficient (ADC) map, areas of restricted diffusion typically appear darker due to the lower ADC values indicating impeded water molecule movement. From your image, there do not seem to be prominent dark areas, which suggests no significant evidence of restricted diffusion in the brain regions shown.

This means there might not be any acute ischemic changes or similar pathologies that significantly restrict water molecule diffusion in these areas of the brain. However, a complete and accurate assessment would require a detailed analysis involving a professional radiologist and correlation with clinical findings.

You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

This Apparent Diffusion Coefficient (ADC) map corresponds to the previously discussed Diffusion Weighted Imaging (DWI) scan. On an ADC map, areas of restricted diffusion typically appear darker compared to surrounding brain tissue because lower ADC values indicate decreased water molecule movement, which is common in acute ischemic strokes and other similar pathologies.

In this image, there is a noticeable dark area in the left frontal region, confirming the evidence of diffusion restriction suggested by the bright area in the DWI scan. This dark appearance in the ADC map supports the diagnosis of restricted diffusion, likely indicating an acute ischemic event or similar pathology in that region of the brain.

This finding should be interpreted alongside clinical symptoms and possibly further diagnostic testing to provide a comprehensive evaluation and appropriate management of the condition.

Figure 5. ChatGPT's incorrect **(a)** and correct **(b)** interpretation of diffusion restriction in acute stroke images. ChatGPT, Chat Generative Pre-trained Transformer.

ChatGPT and other general-purpose LLMs are usually designed to include inherent randomness, which means that their outputs can vary across multiple runs with the same prompt. This feature can enhance user engagement by generating more diverse and dynamic conversations. However, it undermines the precision of GPT-4V when interpreting medical images.

The potential applications of LLMs, such as ChatGPT, in radiology are inspiring.[11] While they can indeed assist radiologists in interpreting images and providing initial assessments, it is crucial to remember that they are not infallible.[14] As with any tool, it has its limitations and can sometimes provide incorrect interpretations. The study by Akinci D'Antonoli et al.[15] likely highlights both the benefits and the challenges of using LLMs in radiology. Although ChatGPT can give false interpretations, it tends to assist experts and give confidence in speeding up certain tasks.[15]

The potential clinical effects of incorporating ChatGPT-4V into radiological practice could be transformative. In settings where radiologists are scarce or imaging interpretation needs to be expedited, ChatGPT-4V could serve as a support tool. This could be particularly impactful in stroke care, where prompt diagnosis is essential.

Although there are existing studies on ChatGPT's role in stroke care, such as "Stroke care in the ChatGPT era: potential use in early symptom recognition" by Lam and Au[16] and "exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study" by Pedro et al.[17], our study is pioneering in evaluating ChatGPT-4V's competence in interpreting stroke images directly. This lack of precedent underscores the novelty and potential significance of our findings in the context of AI-assisted diagnostics.

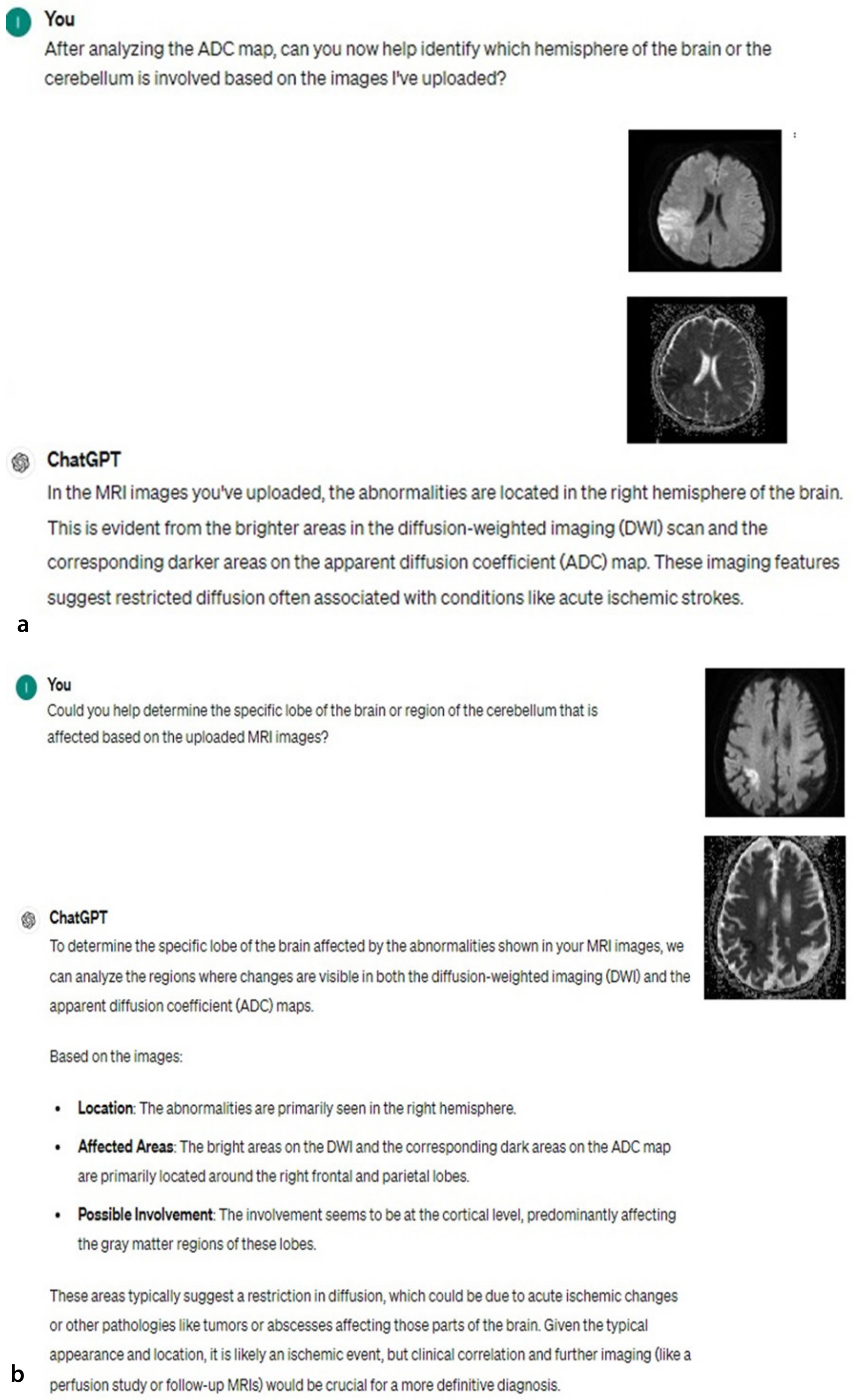
In the Chen et al.[18] study of large vessel occlusion cases, ChatGPT agreed with the physician's decision to perform thrombectomy in 54.3% of cases. ChatGPT had mathematical, logical, and misinterpretation errors in 8.8% of cases. Despite the mistakes, ChatGPT could make nuanced clinical judgments and perform multilevel reasoning.[18] Conversely, the article by Saenger et al.[19] highlighted the diagnostic delay and error caused by misinterpretation from ChatGPT. The patient, who had consulted ChatGPT about his symptoms, had made an underestimation and did not apply to a healthcare institution. As the symptoms progressed, the patient was admitted to the hospital and diagnosed with a TIA. The author reported that this resulted in a serious treatment delay and a potentially life-threatening situation. It was emphasized that with the widespread use of AI, attention should be drawn to such risks, and the final say in the medical decision-making process should belong to healthcare professionals.[19]

Notably, ChatGPT-4V demonstrated a higher success rate in interpreting abnormalities in the frontal and parietal lobes compared with the temporal and occipital lobes. This variation in success may be attributed to the distinctiveness of imaging features or the complexity of the regions involved, suggesting areas for further model training and improvement.

One challenge highlighted by our study is the inconsistent interpretation capabilities of ChatGPT-4V. While showing promise in certain analytical tasks, its performance varied, suggesting that although AI can augment radiological assessments, it currently cannot replace the nuanced judgment of human experts.

The study also draws attention to the lack of transparency in how ChatGPT-4V arrives at its conclusions, a common limitation in AI technologies known as the "black box" issue. This lack of insight into the decision-making process can be a significant barrier to clinical adoption, as understanding the rationale behind diagnostic recommendations is crucial for trust and reliability.

Despite its diagnostic advantages, ChatGPT is not yet a method that can be used independently in time-sensitive situations, such as stroke. The most appropriate use of ChatGPT is as a diagnostic support algorithm under the supervision of a radiologist. If healthcare practitioners utilize ChatGPT, the results must be verified by the radiologist for complete and accurate interpretation.



You

After analyzing the ADC map, can you now help identify which hemisphere of the brain or the cerebellum is involved based on the images I've uploaded?

ChatGPT

In the MRI images you've uploaded, the abnormalities are located in the right hemisphere of the brain. This is evident from the brighter areas in the diffusion-weighted imaging (DWI) scan and the corresponding darker areas on the apparent diffusion coefficient (ADC) map. These imaging features suggest restricted diffusion often associated with conditions like acute ischemic strokes.

You

Could you help determine the specific lobe of the brain or region of the cerebellum that is affected based on the uploaded MRI images?

ChatGPT

To determine the specific lobe of the brain affected by the abnormalities shown in your MRI images, we can analyze the regions where changes are visible in both the diffusion-weighted imaging (DWI) and the apparent diffusion coefficient (ADC) maps.

Based on the images:

- **Location**: The abnormalities are primarily seen in the right hemisphere.
- **Affected Areas**: The bright areas on the DWI and the corresponding dark areas on the ADC map are primarily located around the right frontal and parietal lobes.
- **Possible Involvement**: The involvement seems to be at the cortical level, predominantly affecting the gray matter regions of these lobes.

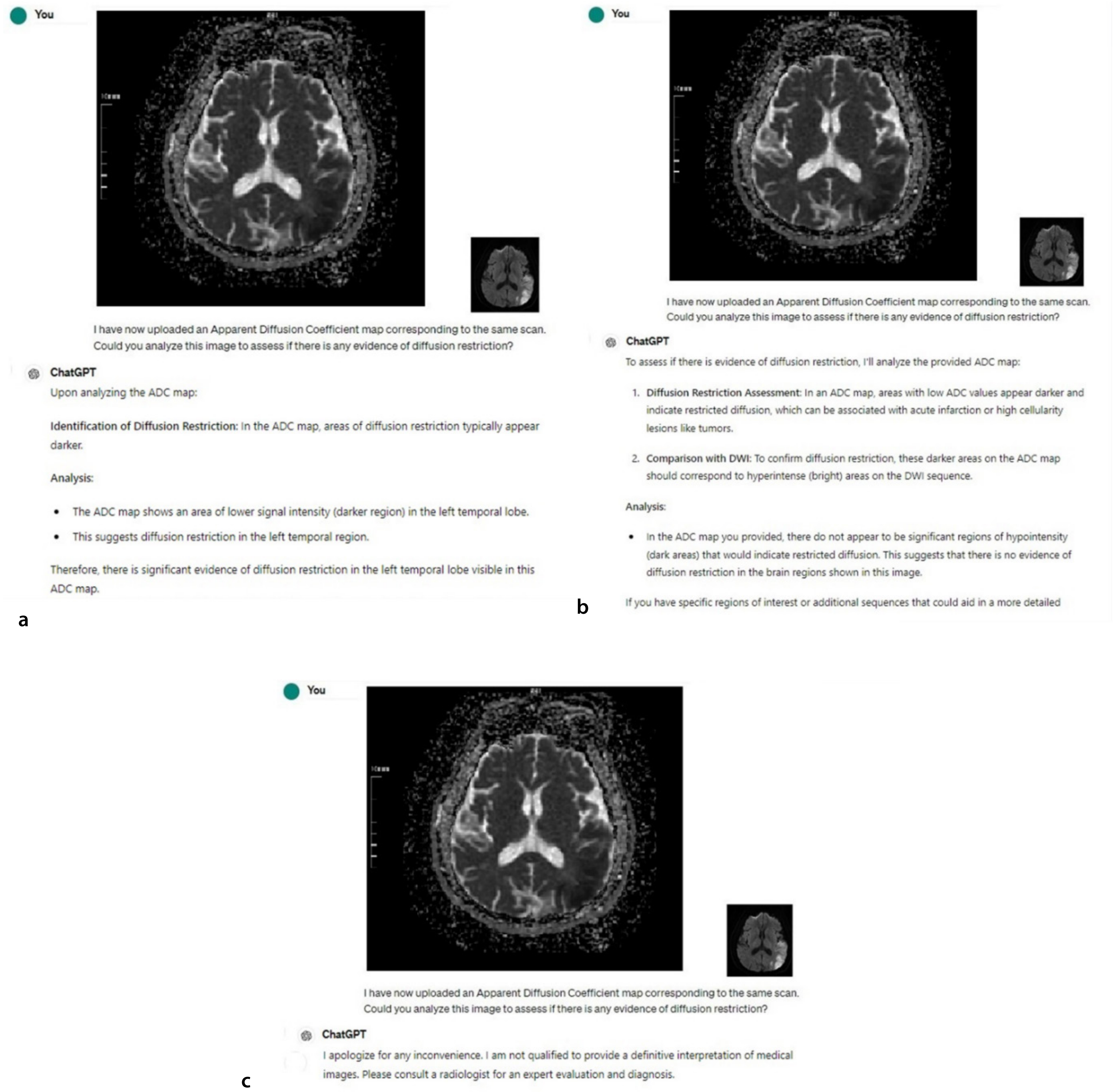
These areas typically suggest a restriction in diffusion, which could be due to acute ischemic changes or other pathologies like tumors or abscesses affecting those parts of the brain. Given the typical appearance and location, it is likely an ischemic event, but clinical correlation and further imaging (like a perfusion study or follow-up MRIs) would be crucial for a more definitive diagnosis.

Figure 6. ChatGPT's correct interpretations (**a, b**) about the location of diffusion restriction in acute stroke images. ChatGPT, Chat Generative Pre-trained Transformer.

Table 2. Diagnostic performance of diffusion restriction test in stroke and normal patients

	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	Accuracy
Number					187/235	202/238	187/223	202/250	389/473
Percentage	187	202	36	48	79.57%	84.87%	83.86%	80.80%	82.24%
95% CI					73.85%–84.54%	79.68%–89.18%	79.24%–87.60%	76.48%–84.49%	78.49%–85.58%

TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.



You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

Upon analyzing the ADC map:

Identification of Diffusion Restriction: In the ADC map, areas of diffusion restriction typically appear darker.

Analysis:

- The ADC map shows an area of lower signal intensity (darker region) in the left temporal lobe.
- This suggests diffusion restriction in the left temporal region.

Therefore, there is significant evidence of diffusion restriction in the left temporal lobe visible in this ADC map.

You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

To assess if there is evidence of diffusion restriction, I'll analyze the provided ADC map:

1. **Diffusion Restriction Assessment:** In an ADC map, areas with low ADC values appear darker and indicate restricted diffusion, which can be associated with acute infarction or high cellularity lesions like tumors.
2. **Comparison with DWI:** To confirm diffusion restriction, these darker areas on the ADC map should correspond to hyperintense (bright) areas on the DWI sequence.

Analysis:

- In the ADC map you provided, there do not appear to be significant regions of hypointensity (dark areas) that would indicate restricted diffusion. This suggests that there is no evidence of diffusion restriction in the brain regions shown in this image.

If you have specific regions of interest or additional sequences that could aid in a more detailed

You

I have now uploaded an Apparent Diffusion Coefficient map corresponding to the same scan. Could you analyze this image to assess if there is any evidence of diffusion restriction?

ChatGPT

I apologize for any inconvenience. I am not qualified to provide a definitive interpretation of medical images. Please consult a radiologist for an expert evaluation and diagnosis.

Figure 7. ChatGPT's correct **(a)** and incorrect **(b)** interpretation and non-response **(c)** about diffusion restriction in acute stroke images. ChatGPT, Chat Generative Pre-trained Transformer.

The limitations of our study include its retrospective design, the potential for selection bias in the images used, and the reliance on a single AI tool for analysis. The evaluation of ChatGPT's performance by a single radiologist presents certain limitations, particularly given the potential for ChatGPT to provide partial or multiple answers. Additionally, not including lacunar infarcts in the study due to diagnostic difficulties may have limited the number of patients. These factors may affect the generalizability of our findings. Future studies should aim to expand the dataset, include prospective analyses, and compare the performance of ChatGPT-4V with other AI models and diagnostic tools. Investigating the integration of AI tools into clinical workflows and their impact on patient outcomes would also be valuable.

In conclusion, despite the current limitations, ChatGPT is a tool with the potential to assist the radiologist in stroke cases where diagnosis timing is very important.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Al-Amin Md, Ali MS, Salam A, et al. History of generative artificial ıntelligence (AI) chatbots: past, present, and future development. 2024. [CrossRef]

2. Adamopoulou E, Moussiades L. An Overview of chatbot technology. 2020;373. [CrossRef]
3. Lee LIT, Kanthasamy S, Ayyalaraju RS, Ganatra R. The current state of artificial intelligence in medical imaging and nuclear medicine. *BJR Open*. 2019;1(1):20190037. [CrossRef]
4. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760. [CrossRef]
5. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering (Basel)*. 2023;10(12):1435. [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. Wang DZ. Editorial comment-- telemedicine: the solution to provide rural stroke coverage and the answer to the shortage of stroke neurologists and radiologists. *Stroke*. 2003;34(12):2957. [CrossRef]
8. Mainali S, Darsie ME, Smetana KS. Machine learning in action: stroke diagnosis and outcome prediction. *Front Neurol*. 2021;12:734345. [CrossRef]
9. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci*. 2023;192(6):3197-3200. [CrossRef]
10. Hou W, Ji Z. GPT-4V exhibits human-like performance in biomedical image classification. *bioRxiv* [Preprint]. 2024:2023.12.31.573796. [CrossRef]
11. Kim H, Kim P, Joo I, Kim JH, Park CM, Yoon SH. ChatGPT vision for radiological interpretation: an investigation using medical school radiology examinations. *Korean J Radiol*. 2024;25(4):403-406. [CrossRef]
12. Deng J, Heybati K, Shammas-Toma M. When vision meets reality: exploring the clinical applicability of GPT-4 with vision. *Clin Imaging*. 2024;108:110101. [CrossRef]
13. Seghier ML. ChatGPT: not all languages are equal. *Nature*. 2023;615(7951):216. [CrossRef]
14. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. *Radiology*. 2023;307(5):e230877. [CrossRef]
15. 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]
16. Lam W, Au SC. Stroke care in the ChatGPT era: potential use in early symptom recognition. *J Acute Dis*. 2023;12(3):129. [CrossRef]
17. Pedro T, Sousa JM, Fonseca L, et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *J Neurointerv Surg*. 2024:jnis-2024-021556. [CrossRef]
18. Chen TC, Couldwell MW, Singer J, et al. Assessing the clinical reasoning of ChatGPT for mechanical thrombectomy in patients with stroke. *J Neurointerv Surg*. 2024;16(3):253-260. [CrossRef]
19. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wien Klin Wochenschr*. 2024;136(7-8):236-238. [CrossRef]