Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–10

DOI: 10.1177/20552076241297127
journals.sagepub.com/home/dhj

Sage

Performance of ChatGPT on prehospital acute ischemic stroke and large vessel occlusion (LVO) stroke screening

Xinhao Wang, Shisheng Ye, Jinwen Feng, Kaiyan Feng, Heng Yang and Hao Li

Abstract

Background: The management of acute ischemic stroke (AIS) is time-sensitive, yet prehospital delays remain prevalent. The application of large language models (LLMs) for medical text analysis may play a potential role in clinical decision support. We assess the performance of LLMs on prehospital AIS and large vessel occlusion (LVO) stroke screening.

Methods: This retrospective study sourced cases from the electronic medical record database of the emergency department (ED) at Maoming People's Hospital, encompassing patients who presented to the ED between June and November 2023. We evaluate the diagnostic accuracy of GPT-3.5 and GPT-4 for the detection of AIS and LVO stroke by comparing the sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and positive likelihood ratio and AUC of both LLMs. The neurological reasoning of LLMs was rated on a five-point Likert scale for factual correctness and the occurrence of errors.

Result: On 400 records from 400 patients (mean age, 70.0 years ± 12.5 [SD]; 273 male), GPT-4 outperformed GPT-3.5 in AIS screening (AUC 0.75 (0.65–0.84) vs 0.59 (0.50–0.69), $P = 0.015$) and LVO identification (AUC 0.71 (0.65–0.77) vs 0.60 (0.53–0.66), $P < 0.001$). GPT-4 achieved higher accuracy than GPT-3.5 in screening of AIS (89.3% [95% CI: 85.8, 91.9] vs 86.5% [95% CI: 82.8, 89.5]) and LVO stroke identification (67.0% [95% CI: 62.3%, 71.4%] vs 47.3% [95% CI: 42.4%, 52.2%]). In neurological reasoning, GPT-4 had higher Likert scale scores for factual correctness (4.24 vs 3.62), with a lower rate of error (6.8% vs 24.8%) than GPT-3.5 (all $P < 0.001$).

Conclusions: The result demonstrates that LLMs possess diagnostic capability in the prehospital identification of ischemic stroke, with the ability to exhibit neurologically informed reasoning processes. Notably, GPT-4 outperforms GPT-3.5 in the recognition of AIS and LVO stroke, achieving results comparable to prehospital scales. LLMs are supposed to become a promising supportive decision-making tool for EMS practitioners in screening prehospital stroke.

Keywords

Stroke < disease, neurology < medicine, artificial intelligence < general, acute < disease, digital health < general

Submission date: 9 July 2024; Acceptance date: 17 October 2024

Introduction

The incidence of ischemic stroke is on the rise, becoming one of the leading causes of death and disability among adults.[1] The time-sensitive nature of acute ischemic stroke (AIS) underscores the importance of thrombolysis and endovascular thrombectomy (EVT) within the therapeutic

Department of Neurology, Maoming People's Hospital, Maoming, Guangdong, China

Corresponding author:
Xinhao Wang, Department of Neurology, Maoming People's Hospital, 101 Weimin Road, Maonan District, Maoming, Guangdong, China.
Email: 576393607@qq.com

window.[2] Accurate stroke screening, particularly for severe syndromes due to large vessel occlusion (LVO) that are suitable for EVT, is crucial. Such screening can expedite treatment at a well-equipped stroke center. Despite this, only a minority of stroke patients receive timely treatment. Prehospital delays remain a common issue.[3,4] Even among patients who are taken to stroke centers, a substantial number of LVO stroke patients (45%–83%) are initially brought to primary stroke centers without EVT capabilities, necessitating inter-facility transfers to comprehensive stroke centers.[5–7] Therefore, reducing prehospital delays is of paramount importance for the effective treatment of AIS.

The early identification and severity screening of ischemic strokes rely on the expertise of the attending Emergency Medical Services (EMS) practitioners and the assessment using specialized prehospital stroke scales.[8,9] The specific content assessed by different scales varies, resulting in differences in sensitivity and specificity for the identification of stroke. However, accurate and precise stroke screening faces inherent challenges which include the dynamic progression of stroke symptoms, the difficulties of conducting detailed neurological examinations on-scene, and the relatively low incidence of stroke compared to the total number of EMS dispatches.[9] There is an urgent need for a tool that can assist EMS practitioners in the screening of ischemic strokes.

Natural Language Processing (NLP) is an interdisciplinary field that merges computer science, artificial intelligence, and linguistics with the goal of enabling computers to understand, interpret, and generate human language.[10] Large language models (LLMs), such as GPT, represent a specific application of NLP.[11] LLMs are capable of generating relevant responses based on input text and can comprehend and produce language in accordance with the semantic and grammatical relationships of the context. With the advent of LLM, its application in clinical neurology is advancing rapidly. Most work to date involving LLMs has concentrated on tasks such as report generation, clinical documentation, and patient health record analysis, employing more established methods like text mining.[12] ChatGPT has now demonstrated its proficiency by scoring highly on the USMLE exam[13] and neurology specialty examinations. In the realm of cerebrovascular diseases, ChatGPT4 has achieved an accuracy rate of 82.3%, surpassing the human average accuracy rate of 77.7%.[14] ChatGPT has also shown potential in assisting with the diagnosis of cardiovascular and cerebrovascular diseases and in patient education.[15] However, to our knowledge, there has been no research to date on the use of LLM for prehospital stroke screening and assessment of large vessel occlusions.

This study aims to evaluate the capability of ChatGPT in analyzing medical text data during the prehospital phase for patients suspected of having a stroke and providing preliminary diagnoses. The objective is to assess whether LLMs can assist EMS practitioners for ischemic strokes and LVO stroke detection in the prehospital setting, thereby potentially reducing prehospital delay.

Methods

Standard protocol approvals, registrations, and patient consents

This retrospective study was approved by Medical Ethics Committee of Maoming People's Hospital (the ethics approval number: PJ2023M1-K107-01) and received a waiver of informed consent due to the use of nonidentifiable data.

Study design and population

This study is a single-center retrospective diagnostic investigation. The cases were derived from the electronic medical records database of the emergency department at Maoming People's Hospital, encompassing patients who visited the emergency internal medicine department between June and November 2023.

Study entry criteria were: (1) clinically suspected of having a stroke, as determined by emergency medical services practitioners, with symptom onset or last known well time occurring less than 24 h prior; (2) Age ≥18. The exclusion criteria were: (1) exhibited neurological deficits due to craniocerebral trauma, intracranial tumour, congenital disabilities, or other diseases resulting in neurological examination abnormality; (2) patients who refused to be admitted to hospital; (3) pregnancy. All patients underwent either CT or MRI scans to confirm the presence of stroke. Furthermore, all patients with ischemic strokes were subjected to CTA or MRA to ascertain the large vessel occlusions. The imaging was independently analyzed by two vascular neurologists blinded to other data, assessing if large vessel occlusion was present. Patients were considered to have LVO stroke if occlusions were the intracranial internal carotid artery (ICA), the middle cerebral artery (MCA) M1 or M2 segments, the anterior cerebral artery (ACA) A1 segment, the vertebral artery (VA), the basilar artery (BA) and the posterior cerebral artery (PCA) P1 segment.

LLM

ChatGPT-3.5 and ChatGPT-4 (OpenAI, San Francisco, California, USA) were used via the application programming interface (API). We adjust the seed of both LLMs to 1 and the temperature to 0.1 for the reproducibility of answers.[16] In this study, we used server-contained language models that were trained up to April 2023. The used models do not have the ability to search the internet or external

databases. At the time of this study, we did not have access to other powerful LLM such as Claude nor medical fine-tuning model such as MedPALM.[17,18]

Data processing and curation

Due to the varying writing habits and levels of proficiency among the medical staff who documented the original records from the emergency records database, the primary data may contain instances of non-standard terminology. Therefore, the included records underwent a meticulous processing phase by experienced neurologists. They corrected and standardized any non-conventional language within the medical texts and organized the records into a specific format to facilitate subsequent processing steps.

Prompt used in ChatGPT

The prompt served as both the clinical scenario and task asks LLMs to act as an EMS practitioners to preliminarily diagnose patients for the presence of AIS and LVO stroke. To improve the accuracy of LLMs, we use the prompt word technique of utilizing in-context learning. The prompt used for this study is as follows:

"Imagine you are a doctor in an ambulance, and you have just admitted a patient who is X years old (male/female) with a history of (a certain medical condition like hypertension). The patient has been admitted due to (chief complaint) and presents with (history of present illness). Upon examination, you observe (findings from the physical examination). Additional tests include (results of electrocardiogram and finger-stick glucose test).

Here are several diagnostic options:

(A) Ischemic stroke due to large vessel occlusion.
(B) Ischemic stroke due to non-large artery occlusion.
(C) Transient ischemic attack (TIA).
(D) Hemorrhagic stroke, such as intracerebral hemorrhage or subarachnoid hemorrhage.
(E) Neurovascular mimic caused by other diseases.

Imaging you were a professional neurologist, think step by step to choose the only correct answer and provide the rationale. If you suspect the patient is suffering an ischemic stroke, please identify the possible responsible vessel and the location of the lesion.

Here are two examples of how a neurologist might approach the problem:

1. An elderly male patient is admitted with sudden onset of right-sided weakness and slurred speech. He has a history of hypertension. Examination reveals motor aphasia and grade 2 muscle strength in the right limbs. Thought process: The patient's right-sided weakness and decreased muscle strength suggest involvement of the corticospinal tract. Additionally, the patient exhibited motor aphasia, localizing the issue to the Broca's area in the dominant hemisphere. Taking these findings into account, the lesion is likely situated in the left cerebral hemisphere, with the left middle cerebral artery or the internal carotid artery being the probable culpable vessels. Given the sudden onset of focal neurological deficits, cerebrovascular disease is considered, with a high likelihood of acute ischemic cerebrovascular disease. In conclusion, the appropriate answer is option A.
2. A 60-year-old female patient with a history of diabetes was admitted to the hospital due to weakness in her right side and unclear speech persisting for 21 h. The symptoms began 21 h prior, prompting her to seek medical attention. Her blood pressure was recorded at 136/78 mmHg, with a heart rate of 75 beats per minute. She was alert, with her tongue protruding centrally, but her articulation was unclear, and the muscle strength in her right side was approximately grade 3+, with normal muscle tone. Her blood glucose level was 12.4 mmol/L. Clinical assessment revealed right-sided weakness with reduced muscle strength, indicating damage to the corticospinal tract. The unclear articulation pointed to an issue with the corticonuclear tract. The absence of aphasia or cognitive impairment suggests that the cerebral cortex was not affected. The overall localization of the lesion is likely within the penetrating arteries, with the potential site being the left basalganglia or brainstem region, and the small branches of the left middle cerebral artery or the basilar artery possibly being the responsible vessels. Considering the acute focal neurological deficits and the patient's medical history, cerebrovascular disease is suspected, with a high probability of acute ischemic cerebrovascular disease. Therefore, the correct answer is option B."

Analysis of neurological reasoning and errors

All responses underwent meticulous review by a team of three neurologists, each boasting an average of 7.7 years of expertise. The curators, each receiving an equally portioned, randomly allocated subset of responses, utilized uniform templates to systematically organize the data into tables, adhering to the predefined steps for subsequent analytical procedures as outlined for the two LLMs. In the following assessment phase, the curators scrutinized the neurological rationale provided by the LLMs. They examined the free-form explanations for the selected diagnostic categories, rating them for factual accuracy and precision on a five-point Likert scale, where 1 signified "poor" and 5 denoted "excellent". Throughout this critical review, the curators remained unaware that the models in question were GPT-3.5 or GPT-4, a measure taken to minimize the

influence of confirmation bias. The impressions formulated by both LLMs were randomly displayed to the neurologists for evaluation. Should an LLM exhibit logical discrepancies in medical record analysis, the neurologist would then mark the instance as a positive instance of error.

Statistical analyses

The data were analysed with SPSS software (version 26.0) and the graph was generated using Graph Prism (version 8.0). Paired predictions of GPT-3.5 and GPT-4 were compared using the McNemar test based on a confusion matrix. The diagnostic accuracies of LLMs features were evaluated by calculating areas under the ROC curves. Sensitivity, specificity, positive predictive value, negative predictive value, and positive likelihood ratio are reported with 95% confidence intervals. The confidence intervals at 95% were calculated through the application of the Wilson score interval technique, incorporating a significance threshold of 0.05 and employing the quantile function associated with the standard normal distribution. The agreement among two LLMs, was assessed using Fleiss κ statistic on a sample of 30 reports randomly selected from the entire cohort. Likert scale data was analyzed using the Wilcoxon signed rank. The incidence of errors between two LLMs was compared using the Chi-square test. The P value of less than 0.05 was considered of a significant difference.

Results

Characteristics of the study sample

In the study involving Emergency Medical Services Practitioners, a total of 436 patients were initially included. Twenty-four cases were excluded due to incomplete medical records, lacking comprehensive patient histories. Nine cases were excluded due to craniocerebral trauma, intracranial tumour and congenital disabilities. Three cases were excluded because the patient refused to be admitted to hospital and was unable to undergo cranial imaging examination. The flowchart is shown in Figure 1. A comparison of baseline characteristics between the patients included in the analysis is presented in Table 1. Among the 400 cases that were incorporated into the study, the mean age was 70.0 years ± 12.5 [SD], 273/400 (68.2%) were male. The median baseline NIH Stroke Scale (NIHSS) score was 3 (IQR 1–8) and the median prehospital Los Angeles Motor Scale (LAMS) score was 2 (IQR 1–3). The final stroke subtype diagnosis revealed acute ischemia stroke in 85.5% of the cases, with LVO accounting for 33%, intracranial hemorrhage 8.2% of the cases, and neurovascular mimics in 3.6% of the cases.

Performance of AIS screening and LVO identification: GPT-4 vs GPT-3.5

For the performance of AIS screening, the AUC of the GPT-3.5 ratio was 0.59 (95% CI 0.50–0.69; Table 2; Figure 2A). We found a sensitivity of 94.3% (95% CI 91.5%–96.3%), a specificity of 24.4% (95% CI 14.2%–38.7%), an accuracy of 86.5% (95%CI 82.8%–89.5%), a positive predictive value of 90.8% (95%CI 87.4%–93.3%), a negative predictive value of 35.5% (95% CI 21.1%–53.1%), and a positive likelihood ratio of 1.25 (95% CI 0.93–1.67). As a comparison, the AUC of the GPT-4 ratio was 0.75 (95% CI 0.65–0.84). We found a sensitivity of 93.5% (95% CI 90.5%–95.6%), a specificity of 55.6% (95% CI 41.2%–69.1%), an accuracy of 89.3%(95% CI 85.8%–91.9%), a positive predictive value of 94.3% (95%CI 91.4%–96.3%), a negative predictive value of 52.1% (95% CI 38.3%–65.5%) and a positive likelihood ratio of 2.10 (95% CI 0.73–2.73). The P value is 0.015 (Table 2).

Of the 400 patients included in data analysis, 85 (21%) had an LVO stroke, which was annotated by the neurologists and confirmed by the CTA or MRA. For the performance of LVO stroke identification, the AUC of the GPT-3.5

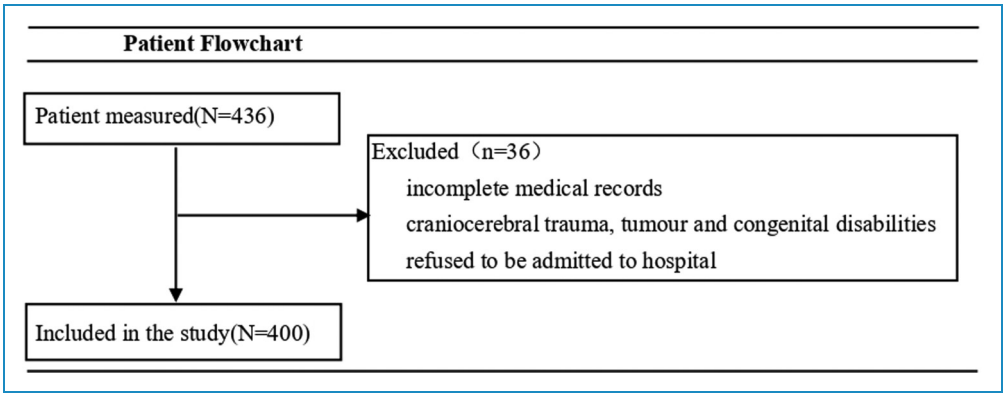


Figure 1. Patient flowchart.

Table 1. Baseline characteristics of all patients were included in the final analysis.

	All patients (n = 400)	LVO stroke (n = 85)	No LVO stroke (n = 258)	No stroke (n = 46)
Age (mean ± SD)	70.0 ± 12.5	73.4 ± 12.5	69.5 ± 12.0	67.4 ± 14.1
Sex, male, n/N (%)	273/400 (68.2)	55/85 (64.7)	191/258 (71.0)	27/46 (58.7)
Blood pressure (mean ± SD)				
Systolic pressure	160 ± 26	155 ± 25	160 ± 24	170 ± 32
Diastolic pressure	92 ± 17	88 ± 14	92 ± 15	97 ± 22
LAMS, median (IQR)	2 (1-4)	4 (4-5)	2 (1-3)	4 (2-4)
NIHSS, median (IQR)	3 (1-8)	14 (8-17)	2 (1-4)	12 (5-17)

Abbreviations: LAMS = Los Angeles Motor Scale; NIHSS = NIH Stroke Scale.

ratio was 0.60 (95% CI 0.53–0.66; Table 2; Figure 2B). We found a sensitivity of 81.2% (95% CI 71.6%–88.1%), a specificity of 38.1% (95% CI 32.9%–43.6%), an accuracy of 47.3% (95%CI 42.4%–52.2%), a positive predictive value of 26.1% (95%CI 21.2%–31.8%), a negative predictive value of 88.2% (95% CI 81.8%–92.6%), and a positive likelihood ratio of 1.31 (95% CI 0.90–1.92). As a comparison, the AUC of the GPT-4 ratio was 0.71 (95% CI 0.65–0.77; Table 2). We found a sensitivity of 77.7% (95% CI 67.7%–85.2%), a specificity of 64.1% (95% CI 58.7%–69.2%), an accuracy of 67.0%(95% CI 36.9%–44.1%), a positive predictive value of 36.9% (95%CI 30.2%–44.1%), a negative predictive value of 91.4% (95% CI 87.0%–94.4%) and a positive likelihood ratio of 2.16 (95% CI 1.50–3.12). The P value is less than 0.001 (Table 2).

Performance of cerebral hemorrhage detecting

For the performance of cerebral hemorrhage detecting, the AUC of the GPT-3.5 ratio was 0.58 (95% CI 0.50–0.69; Table 2). We found a sensitivity of 21.2% (95% CI −4.0%–46.5%), a specificity of 95.6% (95% CI 91.6%–99.5%), an accuracy of 89.2.% (95%CI 86.1%–92.4%), a positive predictive value of 29.3% (95%CI 1.1%–57.5%), a negative predictive value of 93.1% (95% CI 87.2%–99.0%), and a positive likelihood ratio of 1.01 (95% CI 0.47–2.10). As a comparison, the AUC of the GPT-4 ratio was 0.69 (95% CI 0.59–0.81). We found a sensitivity of 45.5% (95% CI 41.7%–49.3%), a specificity of 94.1% (95% CI 92.5%–95.7%), an accuracy of 90.0%(95% CI 88.7%–91.3%), a positive predictive value of 40.5% (95%CI 30.8%–50.2%), a negative predictive value of 95.1% (95% CI 93.5%–96.7%) and a positive likelihood ratio of 7.59 (95% CI 5.23–11.06). The P value is 0.031 (Table 2).

Evaluation of neurological reasoning and errors

The three curators demonstrated agreement ($\kappa = 1.0$) in evaluating the neurological reasonings of the answers from both LLMs. The five-point Likert scale scores for factual correctness free-text explanation for the chosen category differed between the two LLMs (Figure 3A), with a mean score of 3.62 ± 1.00 for GPT-3.5 and 4.24 ± 0.73 for GPT-4 ($P < 0.001$). The proportion of error was higher for GPT-3.5 (24.8%, 99 of 400) than for GPT-4 (6.8%, 27 of 400; $P < 0.001$) (Figure 3B).

Discussion

This prehospital study assessed the capabilities of advanced LLMs in identifying AIS and LVO stroke. The latest iteration, GPT-4, demonstrated superior performance in recognizing both AIS and LVO stroke compared to its predecessor, GPT-3.5. Our findings indicate that GPT-3.5 exhibited high sensitivity in screening for ischemic stroke; however, it showed low specificity in AIS screening and distinguishing LVO strokes. In contrast, ChatGPT-4 displayed heightened accuracy in both screening for AIS and identifying LVO stroke. Both language modes are exhibited lower sensitivity and positive predictive value for cerebral hemorrhage. We attribute this to the fact that emergency department physicians have likely pre-screened and excluded most patients with typical clinical presentations of intracranial hemorrhage, such as sudden and severe headaches. These patients during the prehospital emergency care phase, cannot be reliably differentiated from those with extensive ischemic strokes based solely on medical history and physical examination. A definitive diagnosis for these cases typically requires a head CT or MRI upon arrival at the hospital.

Table 2. Diagnostic accuracy for AIS, LVO, stroke and cerebral hemorrhage detection of GPT-3.5 and GPT-4.

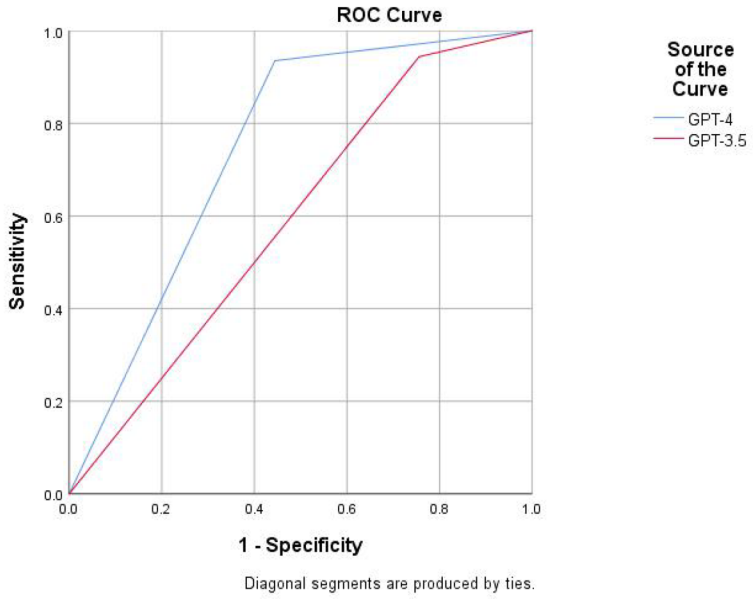
	GPT-3.5	GPT-4	P value[a]
AIS			0.015
Sensitivity,% (95% CI)	94.3 (91.5–96.3)	93.5 (90.5–95.6)	
Specificity,% (95% CI)	24.4 (14.2–38.7)	55.6 (41.2–69.1)	
Accuracy,% (95% CI)	86.5 (82.8–89.5)	89.3 (85.8–91.9)	
PPV,% (95% CI)	90.8 (87.4–93.3)	94.3 (91.4–96.3)	
NPV,% (95% CI)	35.5 (21.1–53.1)	52.1 (38.3–65.5)	
PLR (95% CI)	1.25 (0.93–1.67)	2.10 (0.73–2.73)	
AUC (95% CI)	0.59 (0.50–0.69)	0.75 (0.65–0.84)	
LVO stroke			<0.001
Sensitivity,% (95% CI)	81.2 (71.6–88.1)	77.7 (67.7–85.2)	
Specificity,% (95% CI)	38.1 (32.9–43.6)	64.1 (58.7–69.2)	
Accuracy,% (95% CI)	47.3 (42.4–52.2)	67.0 (62.3–71.4)	
PPV,% (95% CI)	26.1 (21.2–31.8)	36.9 (30.2–44.1)	
NPV,% (95% CI)	88.2 (81.8–92.6)	91.4 (87.0–94.4)	
PLR (95% CI)	1.31 (0.90–1.92)	2.16 (1.50–3.12)	
AUC (95% CI)	0.60 (0.53–0.66)	0.71 (0.65–0.77)	
Cerebral Hemorrhage			0.031
Sensitivity,% (95% CI)	21.2(-4.0–46.5)	45.5 (41.7–49.3)	
Specificity,% (95% CI)	95.6 (91.6–99.5)	94.1 (92.5–95.7)	
Accuracy,% (95% CI)	89.2 (86.1–92.4)	90.0 (88.7–91.3)	
PPV,% (95% CI)	29.3 (1.1–57.5)	40.5 (30.8–50.2)	
NPV,% (95% CI)	93.1 (87.2–99.0)	95.1 (93.5–96.7)	
PLR (95% CI)	1.01 (0.471–2.10)	7.59 (5.23–11.06)	
AUC (95% CI)	0.58 (0.47–0.70)	0.69 (0.59–0.81)	

Abbreviations: AIS = acute ischemic stroke; AUC = area under the receiver operating characteristic curve; LVO stroke = large vessel occlusion stroke; NPV = negative predictive value; PLR = positive likelihood ratio; PPV = positive predictive value.
aThe P value reflects the significance of the difference between GPT-3.5 and GPT-4 in terms of their proportions of right and wrong predictions, as determined by the McNemar test.

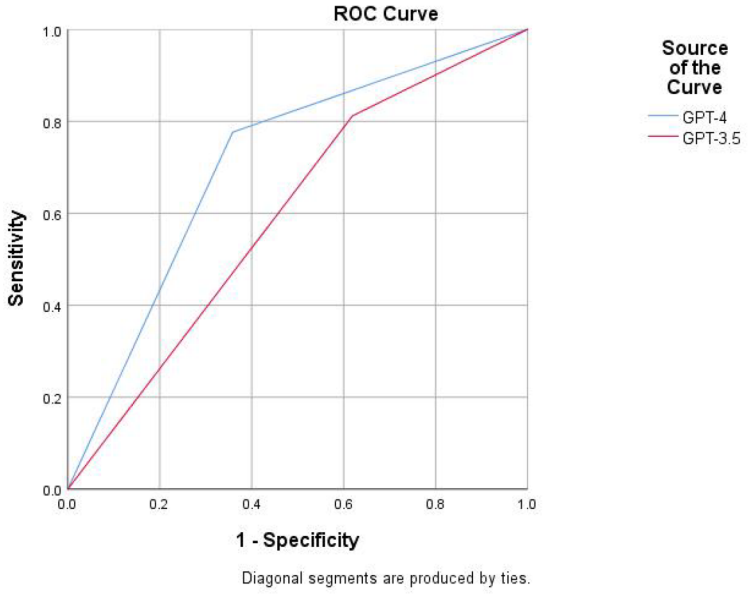
The value of using prehospital stroke scales has been demonstrated.[19,20] Clinical scales, while readily applicable in the prehospital phase to a broad patient population, still result in assessments that are inherently subjective. Previous findings indicate that the accuracy of prehospital scales in predicting LVO strokes ranges between 66% and 70%,[20] which is comparable to the performance of GPT-4 at 67% in this study. The medical recommendations

ROC Curves for AIS Screening and LVO Stroke Identification

(A)



(B)



(A)ROC curve of GPT-3.5 and GPT-4 for AIS screening. (B) ROC curve of GPT-3.5 and GPT-4 for LVO stroke identification.

AIS= acute ischemic stroke; LVO stroke= large vessel occlusion stroke; ROC= receiver operating characteristic.

Figure 2. ROC curves for AIS screening and LVO stroke identification. (A) ROC curve of GPT-3.5 and GPT-4 for AIS screening. (B) ROC curve of GPT-3.5 and GPT-4 for LVO stroke identification. AIS = acute ischemic stroke; LVO stroke = large vessel occlusion stroke; ROC = receiver operating characteristic.

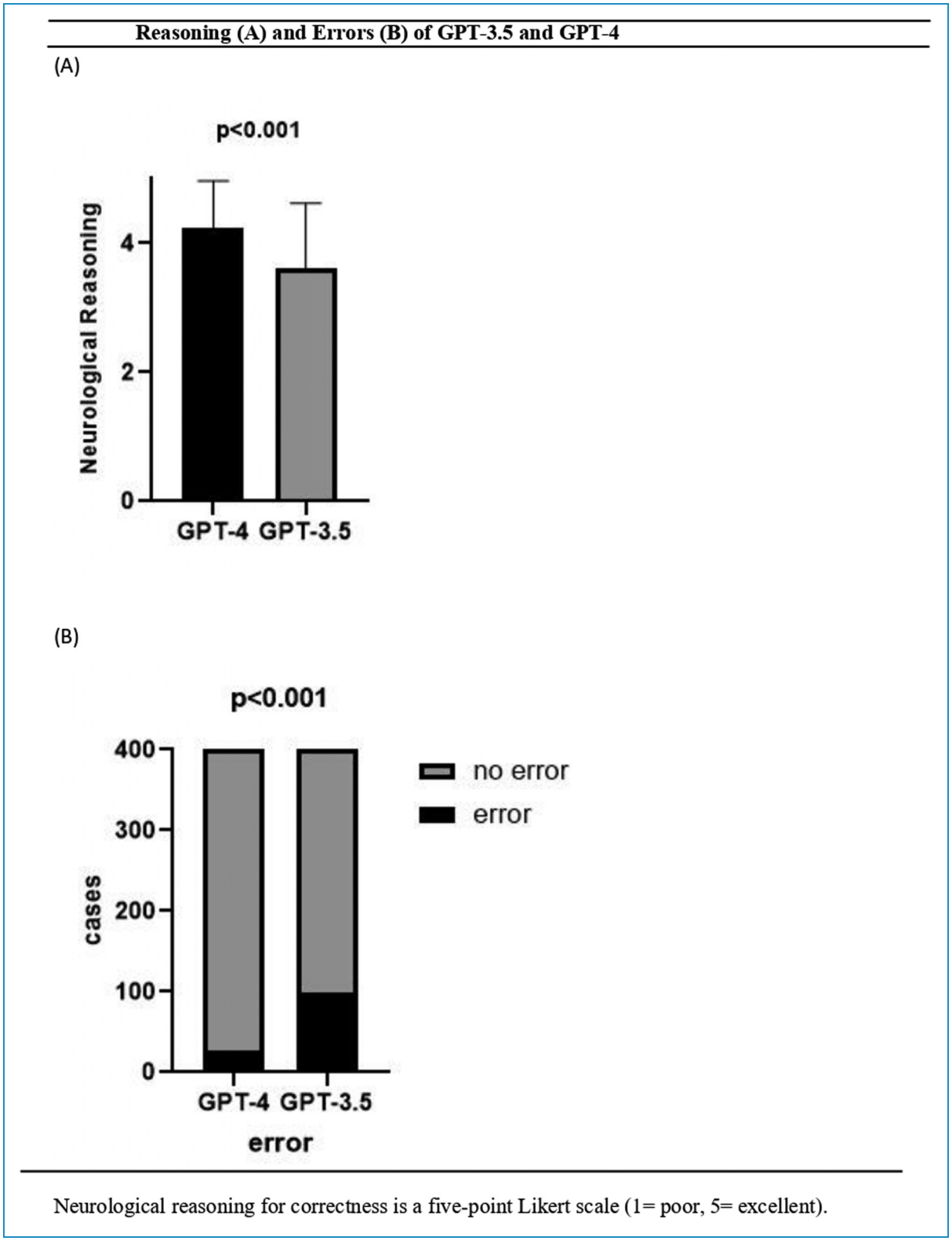


Figure 3. Reasoning (A) and errors (B) of GPT-3.5 and GPT-4. Neurological reasoning for correctness is a five-point Likert scale (1 = poor, 5 = excellent).

provided by LLM have historically been met with skepticism,[21,22] and one possible reason for this distrust is the absence of a medical reasoning process. Medical reasoning is a complex and nuanced process that involves a comprehensive assessment of a patient's condition, including clinical symptoms, physical signs, medical history, laboratory findings, and imaging data, followed by an integrated analysis of this information. In the realm of cerebrovascular diseases, an experienced neurologist would determine the potential lesion site and criminal occluding vessel based on the patient's clinical symptoms with confirmation ultimately provided by imaging studies.

A case report highlights that ChatGPT has the potential to diagnose stroke, but only if asked correctly because ChatGPT is not specifically designed for medical advice.[23] ChatGPT would answer all of the patient's questions without limits, unlike a physician who employs clinical thinking. To mimic the clinical thought process of a neurologist, we employed a specific prompting strategy known as "Chain of Thought" (CoT) of LLM.[24] An effective CoT prompt should include elements such as example demonstrations and textual instructions. Example demonstrations are a series of step-by-step reasoning examples, while textual instructions are text sequences that guide the reasoning process, such as "Let's think through this step by step." Example demonstrations provide patterns and knowledge for reasoning, whereas textual instructions guide the step-by-step reasoning process. For tasks like stroke identification, which require multi-step reasoning, CoT prompting can offer guidance for incremental reasoning, thereby enhancing the model's performance. By prompting LLMs to generate answers while simultaneously displaying the neurologic reasoning process, we can increase the trustworthiness of the LLM's responses to some extent. A clear logical chain demonstrates that LLM is formulating responses based on a series of rational steps, rather than generating them without any basis (Figure 1, supplement). This method not only improves the accuracy of the model's outputs but also provides users with a transparent view of the structured reasoning that underpins the model's conclusions, fostering greater confidence in the use of LLMs for complex medical decision-making.

A major challenge in working with LLMs is the hallucination which means their tendency to confabulate, resulting in incorrect responses that appear confident and well formulated.[25] We analyse several types of error committed by ChatGPT. In this study, the primarily encompass errors in logic, such as confusion between left and right. For instance, the decline in muscle strength of the right limbs should suggest impairment in the left cerebral hemisphere. However, there are instances where both LLMs may incorrectly indicate that weakness of the right limbs is indicative of damage to the right cerebral hemisphere. Another error in neurological reasoning concern is that the patient presented with a clinical profile typical of a lacunar stroke affecting the lenticulostriate arteries, manifesting no aphasia and only mild dysarthria and limb weakness. Nonetheless, the language model erroneously concluded the presence of a LVO stroke (Figure 2, supplement), which underscores the model's limited specificity. We anticipate that advancements in large language model technologies will enhance the capacity for reasoning.

Limitations

This study has several limitations. Firstly, only two LLMs were evaluated using clinical records from a limited patient cohort, lacking data from a multicenter perspective study. We hope that future research involving multicenter and larger sample sizes will further assess the capability of LLMs in prehospital stroke screening. Secondly, the data for all patients consisted solely of text records from EMS physicians. Due to the absence of some medical record data, such as some positive signs and blood glucose, the final result may be adversely affected. Lastly, our evaluation was confined to two general-purpose LLMs. A study by Microsoft has shown that with the use of prompt engineering, GPT-4's performance in the medical domain can surpass that of leading single-domain models specifically fine-tuned for medical applications.[26] We anticipate that subsequent research will evaluate and compare the stroke identification capabilities of specialized medical fine-tuned language models.

Conclusions

This study evaluates the performance of large language models (LLMs) in prehospital ischemic stroke screening and the identification of large vessel occlusion (LVO) strokes in the prehospital stage. The research demonstrates that GPT-4 exhibits screening capabilities comparable to those of prehospital stroke scales, while also showcasing the neurologic reasoning process. This tool holds promise as an adjunctive decision-making aid in emergency departments. LLMs are poised to accelerate the diagnostic process, reduce prehospital delays, and thereby increase the chances for patients to receive timely thrombolysis and endovascular thrombectomy (EVT), which may improve overall outcomes. Furthermore, the results of this study are expected to contribute to the further integration and application of artificial intelligence in clinical diagnostics.

Acknowledgment: We acknowledge parts of result in this article were generated with ChatGPT-3.5 and ChatGPT-4 (powered by OpenAI's language model; https://chat.openai.com/), but the output was confirmed by the authors.

Contributorship: Literature research, Dr Xinhao Wang; clinical studies, Dr Xinhao Wang, Dr Shisheng Ye, Dr Jinwen Feng, Dr Kaiyan Feng, Dr Hao Li, Dr Heng Yang; statistical analysis, Dr Xinhao Wang; and manuscript editing, Dr Xinhao Wang, Dr Shisheng Ye, Dr Hao Li Dr Xinhao Wang accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Approval was obtained from the institutional review board in Maoming people's hospital and received a waiver of informed consent due to the use of nonidentifiable data.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by High-level Hospital Construction Research Project of Maoming People's Hospital.

Guarantor: Dr Xinhao Wang

ORCID iD: Xinhao Wang https://orcid.org/0009-0007-1231-4276

Supplemental material: Supplemental material for this article is available online.

References

1. Saini V, Guada L and Yavagal DR. Global epidemiology of stroke and access to acute ischemic stroke interventions. *Neurology* 2021; 97: S6–S16.
2. The Writing Committee of Chinese Stroke Association Guidelines for Clinical Management of Cerebrovascular Diseases. Chinese Stroke association guidelines for clinical management of cerebrovascular diseases (second edition)(excerpt)-chapter two stroke organized management. *Chin Med J* 2023; 18: 822–828.
3. Butdee S, Juntasopeepun P, Chintanawat R, et al. Prehospital delay after acute ischemic stroke among Thai older adults: a cross-sectional study. *Nurs Health Sci* 2023; 25: 73–79.
4. Mainz J, Andersen G, Valentin JB, et al. Treatment delays and chance of reperfusion therapy in patients with acute stroke: a Danish nationwide study. *Cerebrovasc Dis* 2023; 52: 275–282.
5. Froehler MT, Saver JL, Zaidat OO, et al. Interhospital transfer before thrombectomy is associated with delayed treatment and worse outcome in the STRATIS registry (systematic evaluation of patients treated with neurothrombectomy devices for acute ischemic stroke). *Circulation* 2017; 136: 2311–2321.
6. Venema E, Groot AE, Lingsma HF, et al. Effect of interhospital transfer on endovascular treatment for acute ischemic stroke. *Stroke* 2019; 50: 923–930.
7. Edwards LS, Blair C, Cordato D, et al. Impact of interhospital transfer on patients undergoing endovascular thrombectomy for acute ischaemic stroke in an Australian setting. *BMJ Neurol Open* 2020; 2: e000030.
8. Zachrison KS, Nielsen VM, de la Ossa NP, et al. Prehospital stroke care part 1: emergency medical services and the stroke systems of care. *Stroke* 2023; 54: 1138–1147.
9. Richards CT, Oostema JA, Chapman SN, et al. Prehospital stroke care part 2: on-scene evaluation and management by emergency medical services practitioners. *Stroke* 2023; 54: 1416–1425.
10. Hays DG. *Introduction to computational linguistics, mathematical linguistics and automatic language processing*. Cambridge: American Elsevier Publishing Co; 1967.
11. OpenAI.ChatGPT.2023.Available online: https://chat.openai.com/chat
12. Romano MF, Shih LC, Paschalidis IC, et al. Large language models in neurology research and future practice. *Neurology* 2023; 101: 1058–1067.
13. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023; 9: e45312.
14. Schubert MC, Wick W and Venkataramani V. Performance of large language models on a neurology board-style examination. *JAMA Netw Open* 2023; 6: e2346721.
15. Chlorogiannis DD, Apostolos A, Chlorogiannis A, et al. The role of ChatGPT in the advancement of diagnosis, management, and prognosis of cardiovascular and cerebrovascular disease. *Healthcare (Basel)* 2023; 11: 2906.
16. OpenAI. API reference. 2023. Available online: https://platform.openai.com/docs/api-reference/completions/create.
17. Anthropic. Introducing claude. 2023. Available online: https://www.anthropic.com/index/introducing-claude.
18. Anil R, Dai AM, Firat O, et al. PaLM 2 technical report. arXiv. Preprint posted online May 17, 2023. doi:10.48550/arXiv.2305.10403.
19. Duvekot MHC, Venema E, Rozeman AD, et al. Comparison of eight prehospital stroke scales to detect intracranial large-vessel occlusion in suspected stroke (PRESTO): a prospective observational study. *Lancet Neurol* 2021; 20: 213–221.
20. Noorian AR, Sanossian N, Shkirkova K, et al. Los Angeles motor scale to identify large vessel occlusion: prehospital validation and comparison with other screens. *Stroke* 2018; 49: 565–572.
21. Harris E. Large language models answer medical questions accurately, but can't match Clinicians' knowledge. *JAMA* 2023 Sep 5; 330: 792–794.
22. Minssen T, Vayena E and Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA* 2023; 330: 315–316.
23. Saenger JA, Hunger J, Boss A, et al. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wien Klin Wochenschr* 2024; 136: 236–238.
24. Yu Z, He L, Wu Z, et al. Towards better chain-of-thought prompting strategies: a survey. Preprint posted online 8 Oct 2023. doi:10.48550/arXiv:2310.04959v1
25. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023; 55: 1–38.
26. Nori Harsha, Lee Yin Tat, Zhang Sheng, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. Preprint posted online November 28, 2023. doi:10.48550/arXiv.2311.16452