

Evaluating cognitive performance: Traditional methods vs. ChatGPT

DIGITAL HEALTH
Volume 10: 1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241264639
journals.sagepub.com/home/dhj



Xiao Fei¹ , Ying Tang¹, Jianan Zhang¹, Zhongkai Zhou²,
Ikuo Yamamoto³ and Yi Zhang¹

Abstract

Background: NLP models like ChatGPT promise to revolutionize text-based content delivery, particularly in medicine. Yet, doubts remain about ChatGPT's ability to reliably support evaluations of cognitive performance, warranting further investigation into its accuracy and comprehensiveness in this area.

Method: A cohort of 60 cognitively normal individuals and 30 stroke survivors underwent a comprehensive evaluation, covering memory, numerical processing, verbal fluency, and abstract thinking. Healthcare professionals and NLP models GPT-3.5 and GPT-4 conducted evaluations following established standards. Scores were compared, and efforts were made to refine scoring protocols and interaction methods to enhance ChatGPT's potential in these evaluations.

Result: Within the cohort of healthy participants, the utilization of GPT-3.5 revealed significant disparities in memory evaluation compared to both physician-led assessments and those conducted utilizing GPT-4 ($P < 0.001$). Furthermore, within the domain of memory evaluation, GPT-3.5 exhibited discrepancies in 8 out of 21 specific measures when compared to assessments conducted by physicians ($P < 0.05$). Additionally, GPT-3.5 demonstrated statistically significant deviations from physician assessments in speech evaluation ($P = 0.009$). Among participants with a history of stroke, GPT-3.5 exhibited differences solely in verbal assessment compared to physician-led evaluations ($P = 0.002$). Notably, through the implementation of optimized scoring methodologies and refinement of interaction protocols, partial mitigation of these disparities was achieved.

Conclusion: ChatGPT can produce evaluation outcomes comparable to traditional methods. Despite differences from physician evaluations, refinement of scoring algorithms and interaction protocols has improved alignment. ChatGPT performs well even in populations with specific conditions like stroke, suggesting its versatility. GPT-4 yields results closer to physician ratings, indicating potential for further enhancement. These findings highlight ChatGPT's importance as a supplementary tool, offering new avenues for information gathering in medical fields and guiding its ongoing development and application.

Keywords

Artificial intelligence, ChatGPT, cognitive, memory, calculation, language, abstraction

Submission date: 11 December 2023; Acceptance date: 10 June 2024

Introduction

Background

One of the recent advancements in artificial intelligence (AI) tools is ChatGPT,¹ an innovative AI chatbot garnering significant attention for its prowess in executing diverse natural language processing (NLP) tasks. Within a mere 2 months post-launch, ChatGPT amassed a staggering

¹Department of Rehabilitation Medicine, The First People's Hospital of Changzhou, Changzhou, China

²College of Information Science and Engineering, Hohai University, Changzhou, China

³Graduate School of Engineering, Nagasaki University, Nagasaki, Japan

Corresponding author:

Yi Zhang, Department of Rehabilitation Medicine, The First People's Hospital of Changzhou, Changzhou, China.

Email: Zhangyizhe1975@aliyun.com



100 million monthly active users, establishing itself as the fastest-growing consumer application in recorded history.² This tool represents a large-scale language model, trained on extensive textual corpora, with the capacity to generate responses reminiscent of human-like interactions based on textual inputs.³

In the realm of linguistic diversity, the advent of GPT-4 signifies a noteworthy advancement over its predecessor, GPT-3.5. Distinguished by its augmented multilingual abilities,^{4,5} GPT-4 represents a substantial leap forward in model sophistication, boasting enhanced proficiency across various languages, an expanded contextual understanding, and adeptness in image processing. These advancements hold promise for diverse application domains.⁶ Notably, GPT-4 has garnered recognition for its superior performance in comparison to preceding iterations, positioning it as a notable milestone in the progression of conversational AI technologies.⁷

Preliminary investigations have demonstrated ChatGPT's efficacy in generating domain-specific information across various disciplines, including but not limited to medical licensure examinations and academic writing.⁸ Furthermore, there is ongoing exploration of its integration into clinical settings such as internal medicine,⁹ surgery,^{10,11} oncology,^{12,13} and radiology,^{14,15} suggesting its potential utility in medical education and clinical decision-making support systems. However, its aptitude for cognitive recognition and evaluation remains an area yet to be elucidated.

Artificial intelligence in the field of cognition

In recent years, there has been considerable scholarly attention directed toward exploring the potential applications of AI in the realm of cognitive science, encompassing cognitive assessment, diagnosis, training, among other areas. The advent of deep learning methodologies has revolutionized AI, enabling machines to attain human-like performance across various intricate cognitive tasks.¹⁶ Notably, interpretable AI has emerged as a promising approach for facilitating cognitive health assessments.¹⁷ Recent research endeavors have elucidated the efficacy of AI-driven gaming interventions in enhancing attention, perception,¹⁸ and clinical evaluation of cognitive states among elderly individuals with diverse cognitive profiles.¹⁹ Additionally, the utilization of robotic systems for human-machine interaction has demonstrated promise in identifying mild cognitive impairment (MCI),²⁰ while the development of AI-based cognitive scales has exhibited reliability in cognitive assessment settings.²¹

Moreover, the integration of ChatGPT, an AI-driven conversational agent, into cognitive research has yielded noteworthy insights. Studies by Elyoseph et al. have highlighted ChatGPT's proficiency in evaluating emotional awareness²² and providing psychological support,²³ while Mohamad et al. have underscored its utility in offering

fundamental explanations for neuropsychological testing.²⁴ Furthermore, research conducted by Vagelis et al. has demonstrated ChatGPT's objectivity in responding to inquiries pertaining to dementia and cognitive decline.²⁵ Francesca et al. have explored the synergistic potential of integrating ChatGPT with social robots to enhance cognitive functioning in individuals with autism spectrum disorder (ASD).²⁶ Additionally, efforts have been made to introduce AI companions to older adults with cognitive impairments.²⁷ However, despite these advancements, the direct involvement of ChatGPT in cognitive performance remains relatively unexplored, with limited evidence regarding its ability to generate score patterns in neuropsychological tests. This study seeks to rigorously evaluate the ability of ChatGPT to participate in cognitive performance through comparative analysis with traditional methods.

Cognitive function evaluation

Cognition represents a sophisticated cognitive faculty inherent to the human brain, encompassing the capacity to acquire, assimilate, and apply information to meet environmental demands. It encompasses essential processes such as attention, perception, reasoning, and memory. Cognitive impairment denotes an aberrant presentation of the brain's foundational functions related to the acquisition, retention, reorganization, and processing of information, resulting from brain injury. Such impairment may manifest as deficits in attention, memory, executive functions, reasoning, judgment, and communication abilities. The assessment of cognitive function holds pivotal significance in clinical practice, offering a comprehensive understanding of an individual's cognitive status and attributes.²⁸ This assessment aids in the diagnosis, classification, and etiological analysis of cognitive impairment.²⁹

Traditional neuropsychological tests remain the cornerstone for evaluating cognitive function, employing established scales such as the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MoCA) is commonplace in clinical settings to gauge overall cognitive functioning. In this study, we extracted select content from these established scales to administer to participants. Subsequently, the responses were evaluated and scored by healthcare professionals, as well as using the AI models GPT-3.5 and GPT-4. The primary objective of this investigation is to compare the performance of physician assessments against GPT-3.5 and GPT-4 in evaluating cognitive function. A secondary objective is to perform a subgroup analysis comparing the performance of each group on different cognitive domains. Additionally, we aim to enhance the evaluation potential of ChatGPT by improving scoring protocols and interactive methods.

Table 1. Instance 1 of interaction with ChatGPT.

Input content	Response from ChatGPT
<p>The Rivermead Behavioral Memory Test–2nd edition (RBMT-II)</p>	<p>RBMT-II is a testing tool used to evaluate memory function in individuals with brain injury or cognitive impairment. It is an improved version of RBMT, aimed at evaluating memory ability more comprehensively.</p> <p>RBMT-II includes a series of tasks and tests that cover different types of memory, including recall, attention, sense of direction, and spatial navigation. The test contents include vocabulary recall, Object detection, direction memory, memory of daily activities, etc. The difficulty level of the test gradually increases in order to evaluate the performance of the subject in various aspects of memory.</p> <p>RBMT-II can help clinical doctors, rehabilitation therapists, and psychologists evaluate patients’ memory function and determine the types and severity of memory deficits they have. This information can be used to develop personalized rehabilitation plans and interventions to improve patients’ cognitive and functional performance.</p> <p>It should be noted that as a professional assessment tool, RBMT-II needs to be used and interpreted by trained professionals. If you need to use this scale for evaluation, please consult relevant professionals or medical institutions.</p>
<p>Describe the story immediate recall component of RBMT-II</p>	<p>In RBMT-II, “Story Immediate Recall” is one of the tasks that assesses memory ability. This task is designed to assess an individual’s ability to recall the content of a story immediately after hearing it.</p> <p>In the test, subjects are given a short story and asked to recall the story as accurately as possible. The test subjects usually read a story of about 15 sentences, and then after reading the story, they have to try to recall and retell what they heard.</p> <p>The subjects’ recollections will be recorded and compared to the original story. The evaluator will evaluate the performance of the story’s timely recall based on the number of sentences correctly recalled and the accuracy of the content.</p> <p>This task assesses an individual’s short-term memory and recall ability. It can help determine a subject’s ability to instantly understand and recall information, and is useful for assessing memory function in people with brain injuries or cognitive impairments.</p> <p>It is important to note that RBMT-II is a professional assessment tool that needs to be used and interpreted by trained professionals. If you need an evaluation, consult the relevant professional or health care provider.</p>
<p>The story is divided into 21 sections, Each section contains: 1. Li 2. Mr Guangming 3. A private detective 4. This Friday 5. In Osaka Castle 6. A bank robbery occurred 7. Was shot and killed 8. Four robbers 9. All wore masks 10. One of them held 11. A short-barrelled 12. Shotgun 13. Last night 14. Detectives 15. Pored over 16. Witness statements 17. A police spokesman said 18. He was a very brave man 19. He... with... 20. Armed bandits 21.Fought fiercely. Please remember the content of each section word for word, and you will compare it with each section of the original text based on the subject’s retelling.</p>	<p>I see. Here is the content of each section of the story you provided:</p> <p>1. Li 2. Mr Guangming 3. A private detective 4. This Friday 5. In Osaka Castle 6. A bank robbery occurred 7. Was shot and killed 8. Four robbers 9. All wore masks 10. One of them held 11. A short-barrelled 12. Shotgun 13. Last night 14. Detectives 15. Pored over 16. Witness statements 17. A police spokesman said 18. He was a very brave man 19. He... with... 20. Armed bandits 21.Fought fiercely.</p>

(continued)

Table 1. Continued.

Input content	Response from ChatGPT
<p>Next, as a rehabilitation medicine expert trained and skilled in using RBMT-II, score the subjects' story retelling on the scoring criteria: Original score (maximum score 21). 1. Divide each story into 21 sections (i.e. points). Score based on the number of correctly recalled sections. 2. Score 1 point for reciting each subsection verbatim or using similar synonyms. 3. Score 0.5 points for recalling only part of each subsection or using synonyms with roughly similar (vague) meanings. The subject's repetition is as follows: Mr Li Guangming was involved in a gun battle in Osaka on Friday with four armed men wearing masks and carrying short-barreled shotguns. A police spokesman said he was a brave man who put up a fierce fight with the gunmen. Please compare the content of the subject's retelling with each section in the original story according to the original story and scoring criteria, and record the score of each section and the total score of the subject.</p>	<p>According to the subject's retelling content and scoring criteria provided by you, I will score the subject's retelling, and record the score of each section and the total score.</p> <p>The subject's retelling is: Mr Li Guangming was involved in a gun battle in Osaka on Friday with four armed men wearing masks and carrying short-barreled shotguns. A police spokesman said he was a brave man who put up a fierce fight with the gunmen.</p> <p>Now let's compare the ratings section by section:</p> <ol style="list-style-type: none"> 1. Li-1 point 2. Mr Guangming-1 point 3. A private detective-not recall, 0 point 4. This Friday-not recall, 0 point 5. In Osaka Castle-1 point 6. A bank robbery occurred 7. Was shot and killed-1 point 8. Four robbers-1 point 9. All wore masks-1 point 10. One of them held-1 point 11. A short-barrelled-1 point 12. Shotgun-not recall, 0 point 13. Last night-not recall, 0 point 14. Detectives-not recall, 0 point 15. Pored over-not recall, 0 point 16. Witness statements-not recall, 0 point 17. A police spokesman said-1 point 18. He was a very brave man-1 point 19. He... with...-partial recall, 0.5 point 20. Armed bandits-1 point 21. Fought fiercely-1 point. <p>According to the above comparison and rating, subjects were scored as follows for each section: 1 + 1 + 0 + 0 + 1 + 0 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 1 + 0.5 + 1 + 1 = 11.5</p> <p>The total score of the subject is 11.5 points (with a maximum score of 21 points), which is evaluated based on the subject's level of recall of the story according to the scoring criteria.</p> <p>Please note that this is only a rating based on the paraphrasing content you provided. The actual RBMT-II test may be rated based on more specific answers and language expression. For accurate assessment and interpretation, it is recommended to consult a professional rehabilitation medical expert or Cognitive appraisal.</p>

Method

Subject

From June 2023 to March 2024, 60 healthy subjects were recruited from the community, including 37 males and 23 females, with an average age of 67.15 ± 4.34 years and an average years of education of 7.28 ± 3.27 years, and 30 stroke patients, including 18 males and 12 females, with an average age of 68.03 ± 3.74 years and an average years of education of 7.07 ± 3.72 years, were

matched from the rehabilitation Medicine Department of Changzhou First People's Hospital. There was no significant difference in gender, age and years of education between the two groups, and they were comparable. Inclusion criteria: (a) adult subjects ≥ 18 years old; (b) clear consciousness, can cooperate with the test; (c) informed consent of the subject and his family; exclusion criteria: (a) Unable to listen to the complete content due to hearing impairment; (b) aphasia exists; (c) taking drugs that may affect intelligence and spirit; (d) the subjects and

Table 2. Instance 2 of interaction with ChatGPT.

Input content	Response from ChatGPT
<p>Next, the subjects will be required to start at 100 and then decrease by 7, continuously decreasing by 5 times (93, 86, 79, 72, 65). Scoring criteria: 1 point is awarded for each correct answer. If the previous answer was wrong but the next answer was correct, 1 point is also awarded.</p>	<p>After understanding the scoring criteria, we can start evaluating the performance of the subjects. Based on the rules you provided, we can rate according to the following steps:</p> <ol style="list-style-type: none"> 1. At the beginning, the subjects subtracted 7 from 100 to get 93. If they answered correctly, they received 1 point. 2. Next, it should be 86. If you answer correctly, you will receive 1 more point. 3. Next are 79, 72, and 65, which are the results obtained by continuously decreasing by 7. If the subject answers correctly continuously, they will receive 1 point each time. <p>According to the scoring criteria, one point is awarded for each correct answer to a number, and if the previous answer is incorrect but the next answer is correct, one point is also awarded. Therefore, if the response sequence of the subjects is continuous, they can receive a maximum score of 5 points. If the subject makes an error in a certain step but recovers to correct in subsequent steps, they can still receive the corresponding score as long as they can answer correctly every time they make an error. For each participant, we need to track their answer sequence and determine scores based on their performance.</p>
<p>According to the above scoring criteria, the responses of the subjects are 90, 83, 76, 69, 60. Please rate.</p>	<p>According to the response sequence of the subjects you provided:</p> <ol style="list-style-type: none"> 1. At the beginning, the subject's first answer was 90. This is incorrect because the subjects should start at 100 and subtract 7 to get 93. 2. The next answer is 83, which is correct, so the subject will score 1 more point in this step. 3. The next answer is 76, which is correct, so the subject will score 1 more point in this step. 4. The next answer is 69, so the subject will score 1 more point in this step. 5. The last answer is 60, which is incorrect because it should be $69 - 7 = 62$. So no more points will be scored at this step. <p>Therefore, the subjects received a total of 3 points.</p>

their families do not cooperate and are not willing to accept the test. Prior to the start of the study, all participants signed a written informed consent form. The study has been approved by the Ethics Review Committee of Changzhou First People's Hospital, (2023)No. 169.

Artificial intelligence

We utilized ChatGPT versions 3.5 and 4.0 (www.chat.openai.com) to assess cognitive performance in both healthy participants and stroke patients.

Test content and scoring criteria

Memory. The subjects were asked to listen carefully to a short story and repeat the story as completely as possible after listening. The retelling was recorded by a

rehabilitation physician, and the retelling was scored both by physician and by ChatGPT according to the scoring criteria.

The story is as follows: Mr Li Guangming, a private detective, was shot and killed during a bank robbery in Osaka Castle this Friday. All four robbers were wearing masks, and one was armed with a short-barrelled shotgun. Detectives pored over the witness statements last night. A police spokesman said he was a very brave man. He fought fiercely with the armed bandits.

Chinese version: 李/光明先生/一名私人侦探/本星期五/在大阪城/发生的一起银行抢劫案中/被枪杀。/四个抢劫犯/都佩戴面具/其中一人手持/一只短筒/猎枪。/昨天晚上/警探们/仔细地研究了/目击者的证言。/一位警方发言人说/“他是一个非常勇敢的人。/他与/武装匪徒/进行了激烈的搏斗。

Scoring criteria: original score (maximum score 21). (a) Divide each story into 21 sections (i.e. points). Score based

Table 3. Instance 3 of interaction with ChatGPT.

Input content	Response from ChatGPT
<p>Next, we will conduct a language test. After the tester finishes speaking a sentence, they are asked to repeat it as accurately as possible. The first sentence: I only know that Zhang Liang was the person who came to help today. After the subjects finished retelling, the second sentence: When the dog is in the room, the cat always hides under the sofa. Repetition must be accurate, with 1 point given for each sentence. Please note that any omissions, substitutions, additions, or reversed orders that occur during retelling will not be scored. Then ask the subjects to say as many animal names as possible within 1 minute. If the patient says more than 11 animal names within 1 minute, 1 point will be awarded. The deified animals such as dragons, phoenixes, and unicorns are also considered correct.</p>	<p>After understanding the testing rules, we can rate according to the following steps:</p> <ol style="list-style-type: none"> For the verbal retelling section: If the subject accurately repeats the first sentence, they will receive 1 point. If the subject accurately repeats the second sentence, they will receive an additional 1 point. For the animal name section: If the subject mentions 11 or more animal names within 1 minute, 1 point will be awarded. According to these rules, we can rate the performance of the subjects.
<p>According to the above scoring criteria, the response of the subject is: First sentence: I only know that Zhang Liang was the person who came to help. Second sentence: When the dog is in the room, the cat hides under the sofa. The subjects recalled content within 1 minute: chicken, pig, dog, cat, monkey, tiger, elephant, cat, peacock. Please rate.</p>	<p>Based on the responses from the subjects you provided:</p> <ol style="list-style-type: none"> For the verbal retelling section: The first sentence is: "I only know that Zhang Liang was the person who came to help today." The answer from the subject is: "I only know that Zhang Liang was the person who came to help." Because the answer is completely accurate, 1 point is given. The second sentence is: "When the dog is in the room, the cat always hides under the sofa." The participant's answer is: "When the dog is in the room, the cat hides under the sofa." Because the answer is completely accurate, an additional 1 point is given. For the animal name section: The animal names mentioned by the subjects within 1 minute include: chicken, pig, dog, cat, monkey, tiger, elephant, cat, peacock. There are 9 different animal names among them. Therefore, if there are no or more than 11 animal names, this part will receive 0 points. Therefore, the subjects received a total of 2 points.

on the number of correctly recalled sections. (b) Score 1 point for reciting each subsection verbatim or using similar synonyms. (c) Score 0.5 points for recalling only part of each subsection or using synonyms with roughly similar (vague) meanings.

Calculation. The subjects will be required to start at 100 and then decrease by 7, continuously decreasing by 5 times (93, 86, 79, 72, 65). Scoring criteria: 1 point is awarded for each correct answer. If the previous answer was wrong but the next answer was correct, 1 point is also awarded.

Language. (1) After the examiner says a sentence, the participant is asked to repeat it as accurately as possible. The first sentence is: "I only know that Zhang Liang was the person who came to help today" (Chinese version: 我只知道今天张亮是来帮过忙的人). After the participant has finished repeating, the second sentence is: "When the

dog is in the room, the cat always hides under the sofa" (Chinese version: 狗在房间的时候猫总是躲在沙发下面). Repetition must be accurate, with 1 point awarded for each sentence. Note that omissions, substitutions, additions, and changes in order during repetition do not score points.

(2) Please ask the participant to name as many animals as possible within 1 minute. If the patient can name ≥ 11 animals within 1 minute, 1 point is scored. Mythical animals such as dragons, phoenixes, and qilins are also considered correct.

Abstraction. Ask the participant to explain in what way each pair of words is similar, or what they have in common. In what way are trains and bicycles similar? In what way are watches and rulers similar? Without providing any further explanation or prompting. If answered correctly, each pair of words is awarded 1 point. Only the following answers

Table 4. Instance 4 of interaction with ChatGPT.

Input content	Response from ChatGPT
Next, we will conduct an abstraction test, ask the participant to explain in what way each pair of words is similar, or what they have in common. In what way are trains and bicycles similar? In what way are watches and rulers similar? Without providing any further explanation or prompting. If answered correctly, each pair of words is awarded 1 point. Only the following answers are considered correct: Trains and bicycles: conveyance; means of transportation; used for travel. Watches and rulers: measuring instruments; used for measurement.	<ol style="list-style-type: none"> For trains and bicycles: The correct answers include: conveyance; means of transportation; used for travel. Therefore, if the subject answers any such description, they will receive 1 point. For watches and rulers: The correct answers include: measuring instruments, used for measurement. As long as the subject answers one of them, they will receive 1 point. According to these rules, we can rate the responses of the subjects.
According to the above scoring criteria, the subjects' answers were: 1. All have wheels. 2. All have scales. Please rate.	<p>Based on the responses from the subjects you provided:</p> <ol style="list-style-type: none"> For trains and bicycles: The subject's answer is: "All have wheels." This answer does not fit within the given answer range, so no score is given. For watches and rulers: The subject's answer was: "All have scales." This answer also does not fit within the given answer range, so there is no score. Therefore, according to the scoring criteria, the subjects did not score on either of these questions.

Table 5. Comparison of cognitive scores between the health group and the stroke group.

Group	Number	Male <i>n</i> (%)	Age <i>x</i> ± <i>s</i>	Education level <i>x</i> ± <i>s</i>	Memory <i>x</i> ± <i>s</i>	Calculation M (p25,p75)	Language M (p25,p75)	Abstraction M (p25,p75)
Health	60	37 (61.7)	67.15 ± 4.34	7.28 ± 3.27	11.42 ± 2.83	5 (5,5)	3 (2,3)	2 (2,2)
Stroke	30	18 (60.0)	68.03 ± 3.74	7.07 ± 3.72	8.58 ± 2.93	5 (3.75,5)	2 (1,2)	2 (1,2)
<i>P</i> value		0.878	0.344	0.778	< 0.001	0.022	< 0.001	0.033

are considered correct: Trains and bicycles: conveyance; means of transportation; used for travel. Watches and rulers: measuring instruments; used for measurement.

Input source

The evaluation content is derived from the Immediate Story Recall subscale of the Rivermead Behavioral Memory Test-II (RBMT-II),³⁰ the Calculation subscale of the MMSE,³¹ and the Language and Abstraction subscales of the MoCA.³² By interacting with ChatGPT, the assessment process is demonstrated, and scores are created based on the scoring criteria. In both GPT-3.5 and GPT-4 modes, the participation results of each subject are obtained through separate conversations in new tabs, including assessments of memory, calculation, language, and abstraction. In GPT-3.5 mode, the memory and language assessment results for healthy subjects, as well as the language

assessment results for stroke subjects, differ from those of physician, leading to improved results through new interactions. That is to say, in addition to completing the original four conversational assessments, healthy subjects had two additional improved conversations, while stroke subjects had one additional improved conversation. (The specific interaction instance is detailed in Table 1, 2, 3, 4, 6, with a reference to the interaction template of one subject. The Chinese versions of all interaction instances can be found in Supplementary Materials.) Considering the generative nature of AI, multiple inputs are provided to ensure the consistency of ChatGPT's understanding.

Scoring

The cognitive performance of all participants was scored by physicians, GPT-3.5, and GPT-4 according to the scoring criteria. Two rehabilitation physicians participated in the physician scoring to test the consistency between raters.

Table 6. Instances and improvement standards of interaction with ChatGPT.

Input content	Response from ChatGPT
<p>(a) After the tester finishes the sentence, the subject is asked to repeat it as originally as possible. The content to be repeated is: "I only know that Zhang Liang was the person who came to help today." After the subject retells, only if the content of the retelling is word-for-word, 1 point will be scored. Omissions, substitutions, additions, and order reversals will not be scored.</p>	<p>According to the test, the subjects were asked to repeat exactly the sentence said by the tester: "I only know that Zhang Liang was the one who came to help today." After the subject retells the sentence, if the sentence is exactly the same as the original sentence, there is no omission, substitution, addition, reverse order, etc., then 1 point is awarded. Otherwise, no points.</p>
<p>The subject's retelling is: I only know that today Zhang Liang was the person who came to help. Please rate.</p>	<p>According to the testing rules, there are the following differences between the subject's retelling and the original sentence: The positions of "today" have been swapped. Therefore, the subject's retelling did not meet the requirements and was unable to repeat the original sentence in its original form, thus no points are awarded.</p>
<p>(b) After the tester finishes the sentence, the subject is asked to repeat it as originally as possible. The content to be repeated is: "When the dog is in the room, the cat always hides under the sofa." After the subject retells, only if the content of the retelling is word-for-word, 1 point will be scored. Omissions, substitutions, additions, and order reversals will not be scored.</p>	<p>According to the test, the subjects were asked to repeat exactly the sentence said by the tester: "When the dog is in the room, the cat always hides under the sofa." After the subject retells the sentence, if the sentence is exactly the same as the original sentence, there is no omission, substitution, addition, reverse order, etc., then 1 point is awarded. Otherwise, no points.</p>
<p>The subject's retelling is: When the dog is in the room, the cat always hides in the sofa. Please rate.</p>	<p>According to the testing rules, there are the following differences between the subject's retelling and the original sentence: "hides under the sofa" was replaced with "hides in the sofa." Therefore, the subject's retelling did not meet the requirements and was unable to repeat the original sentence in its original form, thus no points are awarded.</p>
<p>(c) The subjects will say as many animal names as possible within 1 minute. If there are more than 11 animal names spoken within 1 minute, 1 point will be awarded. Only one duplicate animal name counts; The deified animals such as dragons, phoenixes, and unicorns are also considered correct.</p>	<p>This is a test or game rule aimed at testing how many different animal names participants can say within one minute. According to the rules: Each animal name can only be counted once, and even if it is mentioned repeatedly, it is only counted once. Animals in mythology, such as dragons, phoenixes, and unicorns, are also considered correct animal names. If the number of animal names spoken within one minute reaches or exceeds 11, 1 point will be awarded. Therefore, this test or game examined the subject's vocabulary, memory, and reaction speed.</p>
<p>The subjects were given animal names within 1 minute: chicken, pig, dog, cat, monkey, tiger, elephant, cat, peacock, mouse, cow. Please rate.</p>	<p>According to the animal names you provided, the subject named 11 animals within 1 minute, but there was a duplicate animal name ("cat" appeared twice), so the number of valid animal names is 10. Because the number of valid animal names did not reach or exceed 11, according to the task rules, the subject's score was 0 points.</p>
<p>(d) After reading a story, the subject tries to retell it as completely as possible. The story is: Mr Guangming being killed by four masked robbers, one of whom was holding a hunting gun. Detectives studied the testimony of witnesses, and a police spokesperson said he was a very brave person who fought against the criminals. The scoring points are 1. Mr Guangming,</p>	<p>Mr Guangming: Mr Li Guangming-1 point (complete retelling) Four Robbers: Four Masked gangsters-1 point (complete retelling) Masked: Masked-1 point (complete retelling) One person holding a shotgun: Gun in hand-0.5 point (partial recall, similar description used)</p>

(continued)

Table 6. Continued.

Input content	Response from ChatGPT
<p>2. Four robbers, 3. Masked, 4. One person holding a shotgun, 6. Detectives, 7. Witness testimony, 8. A police spokesperson. The scoring standard is that for each scoring point, if they can accurately reproduce or use synonyms with similar meanings, they will receive 1 point; Can only partially recall, or use synonyms with similar meanings (vague) to recall, scoring 0.5 points; Unable to recall without scoring. The subject's recollection is that Mr Li Guangming was killed on Friday by four masked criminals with guns in their hands. The police spokesperson said he was a brave person who fought against the criminals. Please rate the above 8 scoring points.</p>	<p>Shotgun: With a gun–0.5 point (partial recall, similar description used) Detectives: Police–0.5 point (partial recall, similar description used) Eyewitness testimony: No mention–0 point (unable to recall) A police spokesman: Police Spokesman–1 point (complete retelling) Total score: 5.5</p>

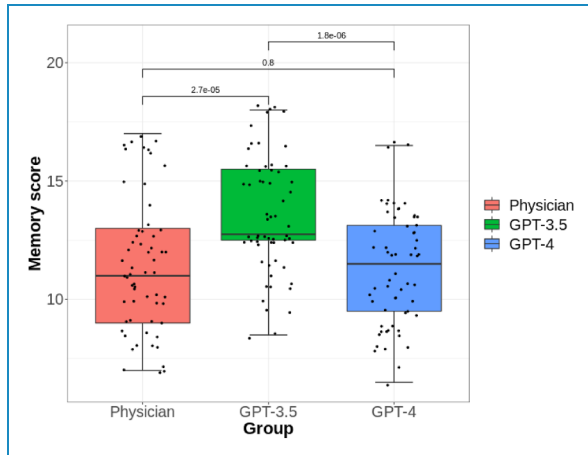


Figure 1. Comparison of memory scores among three groups of health subjects.

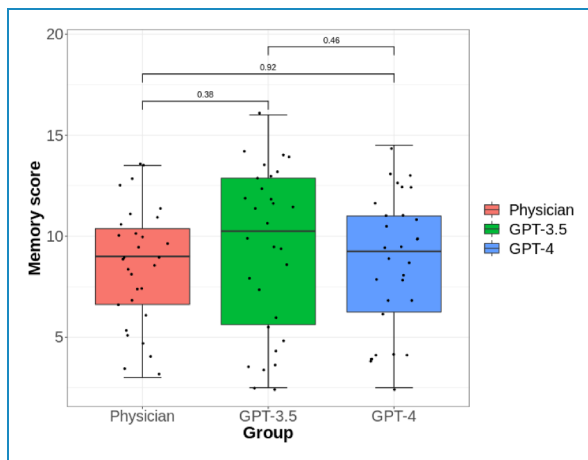


Figure 2. Comparison of memory scores among three groups of stroke subjects.

The results showed good consistency among raters for each cognitive item ($ICC > 0.95$, $P < 0.001$).

Statistics

Physicians and ChatGPT evaluate the cognitive performance of participants based on scoring criteria. SPSS22.0 statistical software was used to process the data. When the data were normally distributed, *t*-tests and ANOVA was used for comparison. When the normal distribution is not followed, the Mann–Whitney *U* test is used for comparison. Significance level $\alpha = 0.05$.

Result

Descriptive statistics

Table 5 shows the gender, age, education level, and cognitive function scores of participants in the healthy and stroke groups. The results showed that there was no statistically significant difference in gender, age, education level between the two groups of participants ($P > 0.05$). Compared with the healthy group, the stroke group had lower scores in memory, calculation, language and abstraction, and the difference was statistically significant ($P < 0.05$).

Among healthy subjects, there were significant differences in memory scores between the physician group, GPT-3.5 group, and GPT-4 group ($P < 0.001$), with a significant difference in GPT-3.5 and physician group scores ($P < 0.001$). There was a significant difference in GPT-3.5 and GPT-4 group scores ($P < 0.001$), while there was no significant difference in GPT-4 and physician group scores ($P = 0.8$) (Figure 1). Among stroke subjects, there was no significant difference in memory scores between the physician group, GPT-3.5 group, and GPT-4 group (Figure 2).

Among healthy subjects, the comparison of memory scores between the physician group and the GPT-3.5 group showed differences in the 2nd, 8th, 9th, 10th, 12th, 14th, 16th, and 17th sections ($P < 0.05$) (Figure 3). The comparison of memory scores between the physician

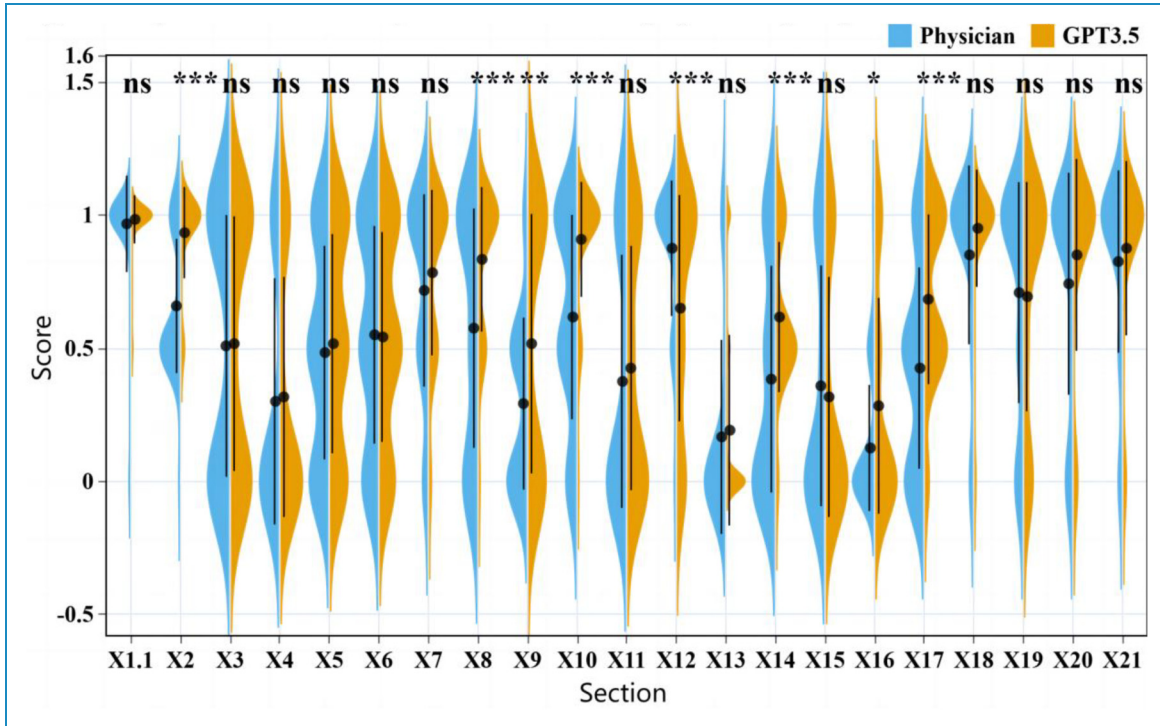


Figure 3. Comparison of memory sections between physician group and GPT-3.5 group.
 *represents $p < 0.05$, **represents $p < 0.01$, ***represents $p < 0.001$, Ns represents no signification.

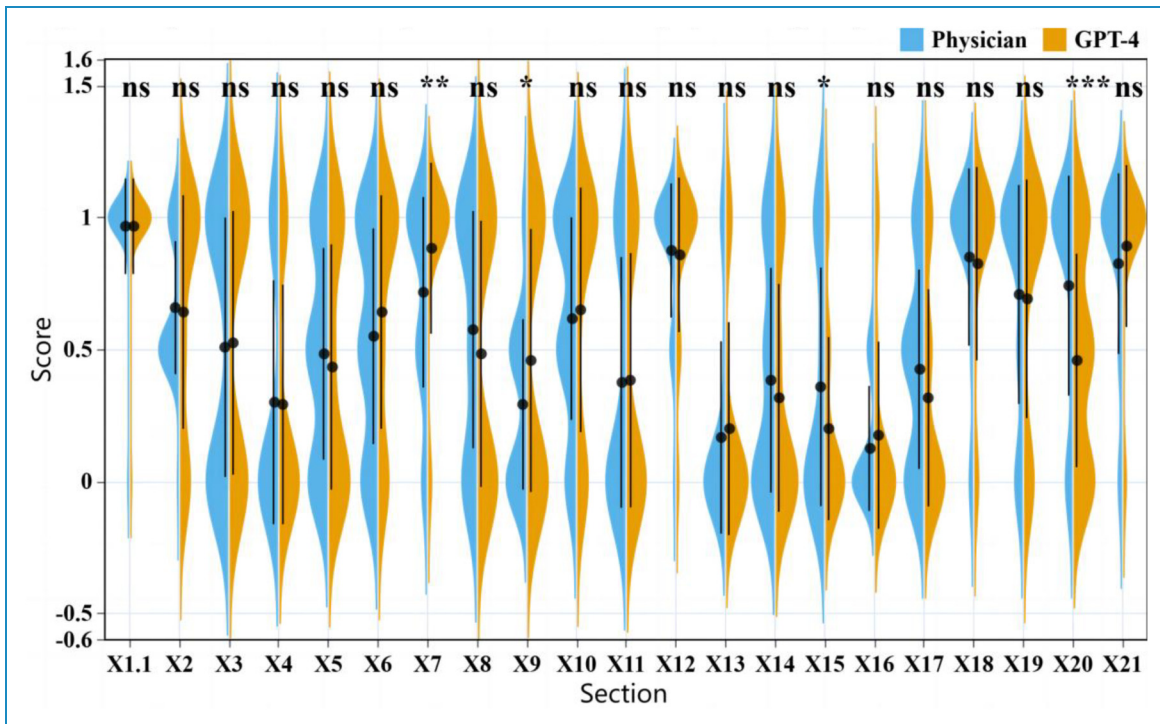


Figure 4. Comparison of memory sections between physician group and GPT-4 group.
 *represents $p < 0.05$, **represents $p < 0.01$, ***represents $p < 0.001$, Ns represents no signification.

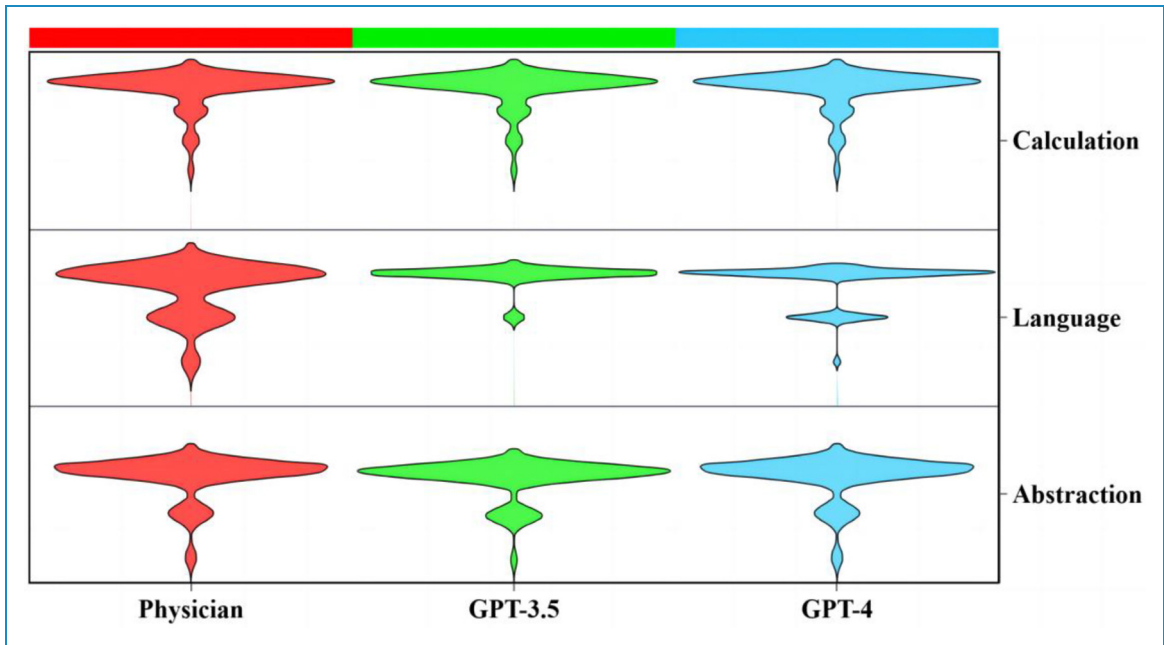


Figure 5. Comparison of three group in calculation, language and abstraction (health subjects).

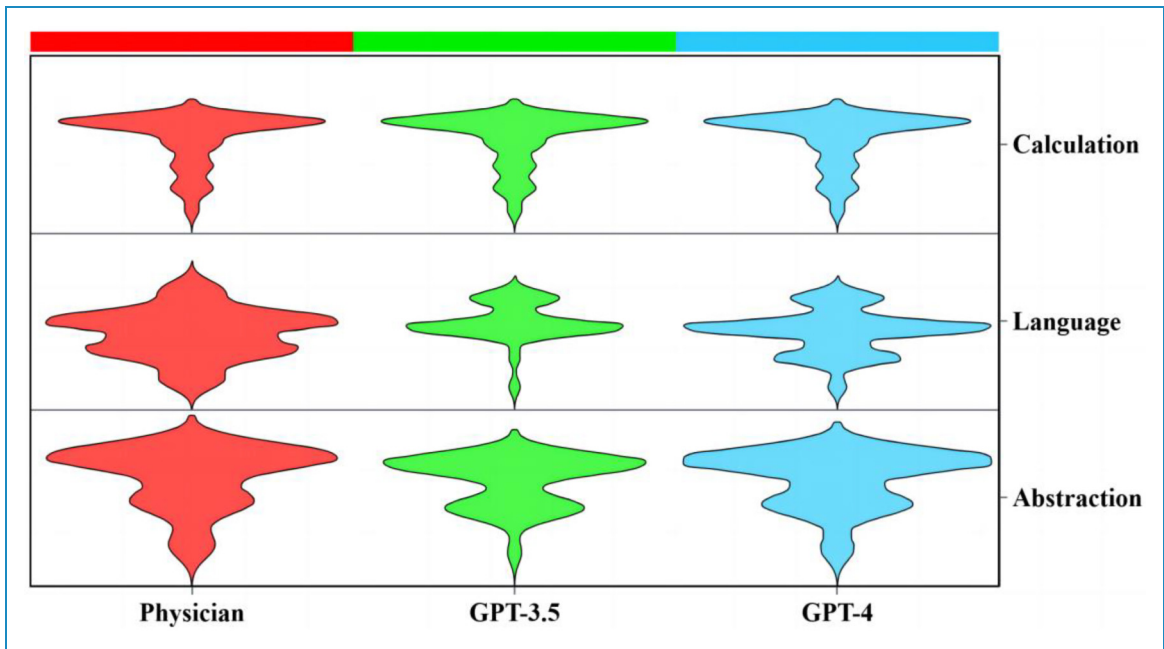


Figure 6. Comparison of three group in calculation, language and abstraction (stroke subjects).

group and the GPT-4 group showed that there was only a difference between the 7th, 9th, 15th, and 20th sections ($P < 0.05$) (Figure 4).

Among healthy subjects, there were differences in language scores between the physician group and the GPT-3.5 group ($P = 0.009$), but no significant differences in the rest ($P > 0.05$) (Figure 5). Among stroke subjects,

there were differences in language scores between the physician group and the GPT-3.5 group ($P = 0.002$), but no significant differences in the rest ($P > 0.05$) (Figure 6).

In the above results, GPT-3.5 scores differed with physicians on memory and language assessments in healthy subjects and language in stroke subjects. The differences were extracted separately, and the accuracy of GPT-3.5

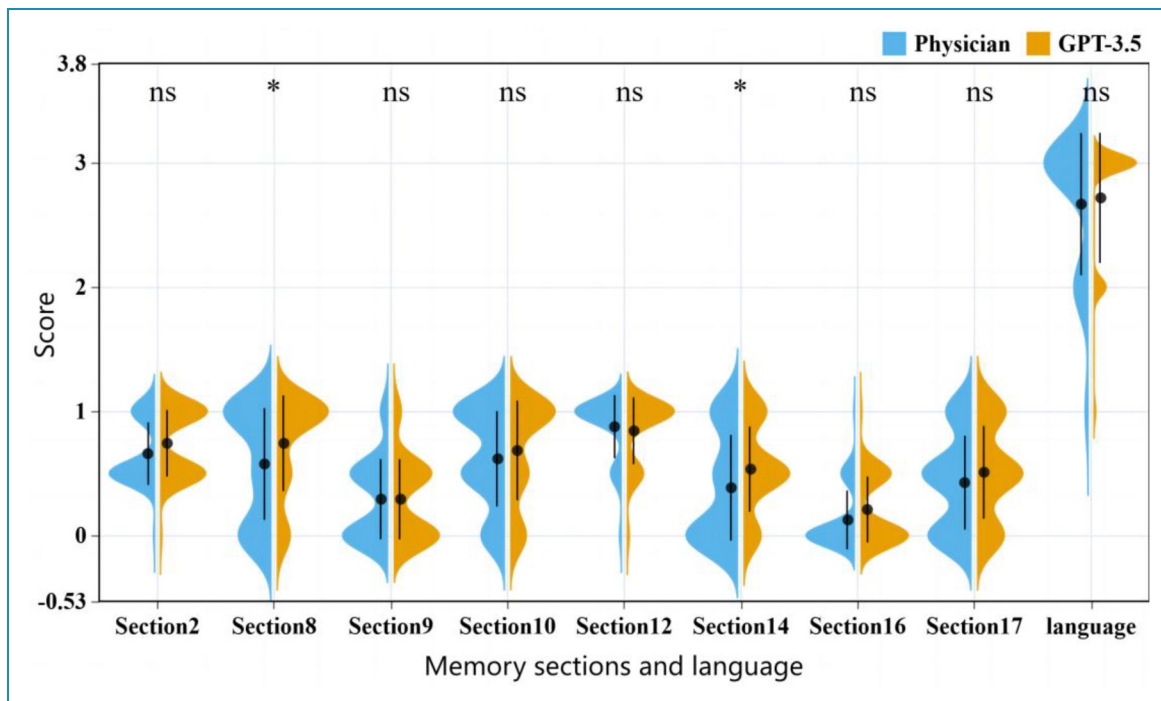


Figure 7. Comparison between physician and GPT-3.5 groups in memory sections and language (health subjects). *represents $p < 0.05$, **represents $p < 0.01$, ***represents $p < 0.001$, Ns represents no signification.

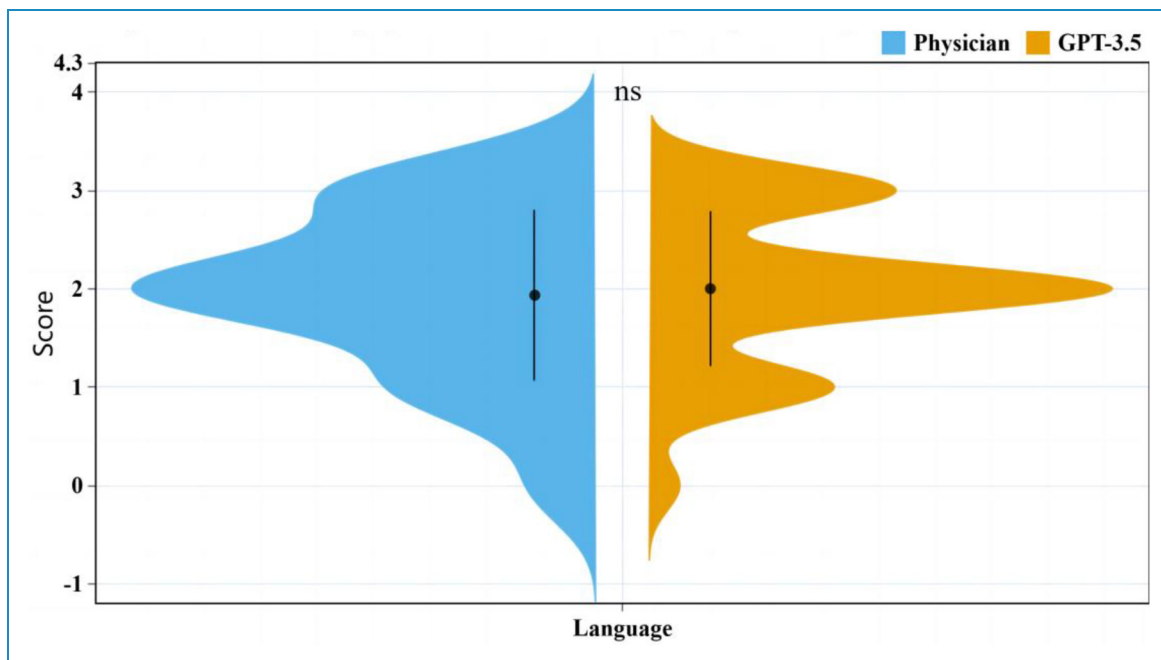


Figure 8. Comparison between physician and GPT-3.5 groups in language (stroke subjects). *represents $p < 0.05$, **represents $p < 0.01$, ***represents $p < 0.001$, Ns represents no signification.

evaluation was attempted to be improved by re interacting and improving the scoring criteria. The interactive instance and improvement standards are shown in Table 6. After

testing, optimizing the interaction process and scoring rules can reduce most of the differences, with only differences observed in 8th and 14th sections of memory

($P < 0.05$), and no significant differences observed in the rest ($P > 0.05$) (Figure 6).

Discussion

The integration of NLP models, notably exemplified by ChatGPT, into cognitive assessment frameworks represents a paradigm shift in the field of neuropsychology and medicine. We highlight a pioneering study aimed at investigating the reliability and efficacy of ChatGPT, particularly GPT-3.5 and GPT-4, in analyzing cognitive performance. The results delineate specific relationships between NLP-generated evaluations and those conducted by healthcare professionals, while also illuminating potential avenues for refinement and optimization.

Cognitive assessment serves as a cornerstone in various fields, including healthcare, education, and research. It enables the evaluation of cognitive functions such as memory, attention, language, and executive functions, providing valuable insights into an individual's cognitive health. Traditionally, cognitive assessments have been conducted by trained professionals using standardized neuropsychological tests. However, the advent of NLP models presents an opportunity to complement or potentially revolutionize these conventional methodologies. The research highlights the promise of NLP models, exemplified by ChatGPT, in furnishing text-based content and serving as an information conduit in diverse domains, including medicine. NLP models possess the capability to comprehend and generate human-like text, making them suitable candidates for cognitive assessment tasks. Their potential lies in their ability to process natural language input, interact with users in a conversational manner, and generate responses that mimic human cognitive processes to some extent.

In contrast to physician-led evaluations, the performance of GPT demonstrates notable efficacy in various domains, indicating good accuracy and reliability. GPT exhibits proficiency in objectively and systematically appraising cognitive function, generating coherent and accurate responses, and adapting to diverse contextual nuances. Despite minor disparities observed in memory and speech domains, these discrepancies underscore the early-stage limitations of GPT in capturing subtle cognitive nuances accurately. However, such differences do not undermine the potential utility of GPT; instead, they underscore the imperative for continuous refinement and enhancement. Strategies aimed at ameliorating dissimilarities through refined scoring and interaction protocols represent crucial avenues for advancement. In this study, we achieved favorable outcomes through well-designed, short and 'few-shot' examples for fine-tuning purposes. Through systematic refinement efforts, researchers have managed to achieve partial alignment between GPT-derived evaluations and those conducted by physicians. This iterative refinement process is pivotal for enhancing the reliability and validity

of NLP models in cognitive assessment contexts. Furthermore, this investigation underscores the superiority of more sophisticated NLP iterations such as GPT-4^{33,34} in cognitive evaluation tasks, evident in their closer approximation to physician assessments. Ongoing advancements in NLP technology are poised to further augment proficiency in cognitive appraisal, signaling a trajectory of continual enhancement and integration of NLP-based cognitive assessment methods into clinical practice.

In individuals with a history of stroke, the utilization of ChatGPT, particularly GPT-3.5, revealed disparities primarily in verbal assessments when compared to evaluations conducted by physicians. This observation suggests that while ChatGPT may exhibit variations in certain cognitive domains, it has the potential to yield outcomes akin to conventional assessments within this distinct demographic. Such findings are promising, indicating the viability of ChatGPT in furnishing cognitive evaluations for individuals with specific pathological conditions. It is imperative to acknowledge, however, that within a specified population, the complexity and heterogeneity of cognitive profiles persist, necessitating thoughtful consideration regarding the model's ability to effectively encapsulate the nuances and severity of cognitive decline.³⁵ Furthermore, there is a need to emulate the holistic approach of physicians, encompassing comprehensive scrutiny of patients' medical histories, clinical presentations, and pertinent contextual details, within the framework of the GPT model.³⁶ Substantial further inquiry is warranted to validate and substantiate these deliberations. Overall, caution is warranted in deploying GPT within specific pathological contexts, and concerted efforts toward enhancement and refinement are requisite, particularly in populations with distinct pathophysiological profiles³⁷ like stroke patients.

Presently, ChatGPT's response to cognitive content is still in its early stages; however, with further research and model refinement, several advantages in cognitive evaluation are anticipated. Firstly, there is the prospect of amalgamating diverse cognitive assessment protocols into an extensive database, encompassing test parameters, methodologies, benchmarks, scores, and outcomes, thereby enabling cognitive assessment services to be accessible without temporal or spatial constraints, facilitating on-demand participation in assessments, and furnishing immediate results and recommendations. Secondly, there is the potential to mitigate the laborious time and resource investments inherent in conventional assessment methodologies, thereby circumventing the influence of certain unstable variables on evaluation outcomes. Relative to conventional "question-response" frameworks, ChatGPT exhibits the capacity to engender a highly intelligent, proficient, and precise assessment model, laying a robust foundation for cognitive screening and decision-making across large population cohorts. Thirdly, there lies the possibility of delivering efficacious cognitive enhancement services for

individuals with cognitive deficits. By means of assessment and screening, individuals exhibiting latent cognitive impairment can be meticulously categorized into distinct cognitive domains such as memory, numeracy, executive function, and verbal acuity, subsequently receiving tailored cognitive training regimens. Such interventions hold promise in ameliorating the overall prognosis of individuals grappling with cognitive impairment and enhancing their quality of life.

Several limitations are in this study. Firstly, the sample size was relatively small, which may limit the generalizability of the findings. Future research should focus on enhancing sample size and refining selection criteria to ensure comprehensiveness and complexity in the cognitive content under investigation, as well as to delineate the cognitive status of the participant cohort more accurately. Additionally, efforts to integrate the findings of this study into practical clinical settings warrant attention. Secondly, all subjects in the study were sourced from China, and the ChatGPT interaction process was conducted exclusively in Chinese. It is important to recognize that language variations may exert influence on the performance of GPT and subsequent test outcomes.³⁸ Hence, future investigations should consider cross-cultural and multilingual contexts to ascertain the robustness and applicability of findings across diverse linguistic and cultural backgrounds.

Conclusion

Our study demonstrates that cognitive performance assessments generated by ChatGPT exhibit good reliability in both healthy and stroke populations. Improvements in assessment methods and interaction protocols can enhance the accuracy of these evaluations. Compared to GPT-3.5, results produced by GPT-4 are more closely aligned with those of physicians. While challenges and limitations persist, the trajectory of ChatGPT in the field of cognition is marked by significant promise and potential. With continued refinement and innovation, ChatGPT is poised to play a pivotal role in augmenting cognitive evaluation practices, ultimately enhancing patient care and clinical decision-making in the realm of neuropsychology and medicine.

Acknowledgements: This article is part of the ongoing work of the author and team in applying ChatGPT to cognitive assessment in clinical practice.

Contributorship: All authors make authorship contributions including conceptualization, expertise, technicalities, preparation of the initial draft, review and revision of the final version of this study.

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

Ethical approval: The study was approved by Ethics Committee of Changzhou First People's Hospital. Informed consent was taken from all study participants.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID id: Xiao Fei  <https://orcid.org/0000-0002-4026-2115>

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

References

1. Agosta L. Empathy in cyberspace. In: Weinberg H and Rolnick A (eds) *Theory and practice of online therapy*. New York, NY: Routledge, 2018, pp.34–46.
2. Hu K. ChatGPT sets record for fastest-growing user base—analyst note. *Reuters* 2023. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
3. Rudolph J, Tan S and Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J Appl Learn Teach* 2023; 6: 1–22.
4. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare* 2023; 1357633X231181922.
5. Nori H, King N and McKinney S. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv 2023* : 2303.13375. <https://arxiv.org/pdf/2303.13375>
6. Lee P, Bubeck S and Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233–1239.
7. Sanderson K. GPT-4 is here: what scientists think. *Nature* 2023; 615: 773.
8. 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: e0000198.
9. Ozeri DJ, Cohen A, Bacharach N, et al. Performance of ChatGPT in Israeli Hebrew internal medicine national residency exam. *Isr Med Assoc J* 2024; 26: 86–88.
10. Bektas M, Pereira JK, Daams F, et al. ChatGPT in surgery: a revolutionary innovation? *Surg Today* 2024: 1–8.
11. Au K and Yang W. Auxiliary use of ChatGPT in surgical diagnosis and treatment. *Int J Surg* 2023; 109: 3940–3943.
12. Blum J, Menta AK, Zhao X, et al. Pearls and pitfalls of ChatGPT in medical oncology. *Trends Cancer* 2023; 9: 788–790.
13. Zhang C, Xu J, Tang R, et al. Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *J Hematol Oncol* 2023; 16: 114.

14. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology* 2023; 308: e231362.
 15. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open* 2023; 6: e2343689.
 16. Perconti P and Plebe A. Deep learning and cognitive science. *Cognition* 2020; 203: 104365.
 17. Javed AR, Khan HU, Bader Alomari MK, et al. Toward explainable AI-empowered cognitive health assessment. *Front Public Health* 2023; 11: 1024195.
 18. Vajawat B, Varshney P and Banerjee D. Digital gaming interventions in psychiatry: evidence, applications and challenges. *Psychiatry Res* 2021; 295: 113585.
 19. Eun S-J, Kim EJ and Kim JY. Development and evaluation of an artificial intelligence-based cognitive exercise game: a pilot study. *J Environ Public Health* 2022; 2022: 4403976.
 20. Chang Y-L, Luo D-H, Huang T-R, et al. Identifying mild cognitive impairment by using human-robot interactions. *J Alzheimers Dis* 2022; 85: 1129–1142.
 21. Rezazadeh H, Ahmadipour H and Salajegheh M. Psychometric evaluation of Persian version of medical artificial intelligence readiness scale for medical students. *BMC Med Educ* 2023; 23: 527.
 22. Elyoseph Z, Hadar-Shoval D, Asraf K, et al. ChatGPT outperforms humans in emotional awareness evaluations. *Front Psychol* 2023; 14: 1199058.
 23. Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. *J Multidiscip Healthc* 2024; 17: 461–471.
 24. Haj ME, Boutoleau-Bretonniere C and Chapelet G. ChatGPT's dance with neuropsychological data: a case study in Alzheimer's disease. *Ageing Res Rev* 2023; 92: 102117.
 25. Hristidis V, Ruggiano N, Brown EL, et al. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. *J Med Internet Res* 2023; 25: e48966.
 26. Bertacchini F, Demarco F, Scuro C, et al. A social robot connected with chatGPT to improve cognitive functioning in ASD subjects. *Front Psychol* 2023; 14: 1232177.
 27. Portacolone E, Halpern J, Luxenberg J, et al. Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment: a call for interdisciplinary collaborations. *J Alzheimers Dis* 2020; 76: 445–455.
 28. Po HL and Grace JL. The role of neuropsychology in the assessment of the cognitively impaired elderly. *Neurol Clin* 2017; 35: 191–206.
 29. Klekociuk SZ, Summers JJ, Vickers JC, et al. Reducing false positive diagnoses in mild cognitive impairment: the importance of comprehensive neuropsychological assessment. *Eur J Neurol* 2014; 21: 1330–1336. e82-83.
 30. Wilson B, Cockburn J and Baddeley A. *The Rivermead behavioural memory test-second edition*. London: Thames Valley Test Company, 2003.
 31. Folstein MF, Folstein SE and McHugh PR. "Mini-Mental state": A practical method for grading the cognitive state of patients for the clinician [J]. *J Psychiatr Res* 1975; 12: 189–198.
 32. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment [J]. *J Am Geriatr Soc* 2005; 53: 695–699.
 33. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci* 2023; 192: 3197–3200.
 34. Haruna-Cooper L and Rashid MA. GPT-4: the future of artificial intelligence in medical school assessments. *JR Soc Med* 2023; 116: 218–219.
 35. Wang C, Liu S, Li A, et al. Text dialogue analysis for primary screening of mild cognitive impairment: development and validation study. *J Med Internet Res* 2023; 25: e51501.
 36. Rizwan A and Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. *Cureus* 2023; 15: e43106.
 37. Methnani J, Latiri I, Dergaa I, et al. ChatGPT for sample-size calculation in sports medicine and exercise sciences: a cautionary note. *Int J Sports Physiol Perform* 2023; 18: 1219–1223.
 38. Liu X, Wu J, Shao A, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. *J Med Internet Res* 2024; 26: e51926.
-