Effectiveness of Large Language Models in Stroke Rehabilitation Health Education: A Comparative Study of ChatGPT-4, MedGo, Qwen, and ERNIE Bot

Shiqi Qiang, Yang Liao, Yanfen Gu, Yanfen Gu, Yue Zhang, Yiyan Wang, Zehui Xu, Zehui Xu, Nuo Han, Haitao Zhang, Haipin Yu

Submitted to: Journal of Medical Internet Research
on: February 28, 2025

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Effectiveness of Large Language Models in Stroke Rehabilitation Health Education: A Comparative Study of ChatGPT-4, MedGo, Qwen, and ERNIE Bot

Shiqi Qiang[1*] MSc; Yang Liao[1*] MSc; Yanfen Gu[2] MSc; Yanfen Gu[2] MSc; Yue Zhang[3] PhD; Yiyan Wang[1] MSc; Zehui Xu[4] BSc; Zehui Xu[4] BSc; Nuo Han[5] BSc; Haitao Zhang[6] MSc; Haipin Yu[7] PhD

[1]Neurorehabilitation Center Shanghai Sunshine Rehabilitation Center Shanghai CN
[2]Department of Gastrointestinal Endoscopy Shanghai East Hospital Shanghai CN
[3]Stroke Center Shanghai East Hospital Shanghai CN
[4]Department of Breast Diseases Yueyang Hospital of Integrated Traditional Chinese and Western Medicine Shanghai University of Traditional Chinese Medicine Shanghai CN
[5] School of Acupuncture-Moxibustion and Tuina Shanghai University of Traditional Chinese Medicine Shanghai CN
[6]Department of Emergency and Critical Care Medicine Shanghai East Hospital Shanghai CN
[7]Department of Nursing Shanghai East Hospital Shanghai CN
[*]these authors contributed equally

Corresponding Author:
Haipin Yu PhD
Department of Nursing
Shanghai East Hospital
YunTai Road No1800
Shanghai
CN

Abstract

Background: Stroke is a leading cause of disability and death worldwide, with home-based rehabilitation playing a crucial role in improving patient prognosis and quality of life. Traditional health education models often fall short in terms of precision, personalization, and accessibility. In contrast, large language models (LLMs) are gaining attention for their potential in medical health education, owing to their advanced natural language processing capabilities. However, the effectiveness of LLMs in home-based stroke rehabilitation remains uncertain.

Objective: This study evaluates the effectiveness of four LLMs—ChatGPT-4, MedGo, Qwen, and ERNIE Bot—in home-based stroke rehabilitation. The aim is to offer stroke patients more precise and secure health education pathways while exploring the feasibility of using LLMs to guide health education.

Methods: In the first phase of this study, a literature review and expert interviews identified 15 common questions and 2 clinical cases relevant to stroke patients in home-based rehabilitation. These were input into four LLMs for simulated consultations. Six medical experts (2 clinicians, 2 nursing specialists, and 2 rehabilitation therapists) evaluated the LLM-generated responses using a Likert 5-point scale, assessing accuracy, completeness, readability, safety, and humanity. In the second phase, the top two performing models from phase one were selected. Thirty stroke patients undergoing home-based rehabilitation were recruited. Each patient asked both models 3 questions, rated the responses using a satisfaction scale, and assessed readability, text length, and recommended reading age using a Chinese readability analysis tool. Data were analyzed using one-way ANOVA, post hoc Tukey HSD tests, and paired t-tests.

Results: The results revealed significant differences across the four models in five dimensions: accuracy ($P = .002$), completeness ($P < .001$), readability ($P = .04$), safety ($P = .007$), and humanity ($P < .001$). ChatGPT-4 outperformed all models in each dimension, with scores for accuracy ($M = 4.28$, $SD = 0.84$), completeness ($M = 4.35$, $SD = 0.75$), readability ($M = 4.28$, $SD = 0.85$), safety ($M = 4.38$, $SD = 0.81$), and user-friendliness ($M = 4.65$, $SD = 0.66$). MedGo excelled in accuracy ($M = 4.06$, $SD = 0.78$) and completeness ($M = 4.06$, $SD = 0.74$). Qwen and ERNIE Bot scored significantly lower across all five dimensions compared to ChatGPT-4 and MedGo. ChatGPT-4 generated the longest responses ($M = 1338.35$, $SD = 236.03$) and had the highest readability score ($M = 12.88$). In the second phase, ChatGPT-4 performed the best overall, while MedGo provided the clearest responses.

Conclusions: LLMs have shown strong performance in home-based stroke rehabilitation education, demonstrating significant potential for real-world applications. However, further improvements are needed in accuracy, professionalism, and oversight.

(JMIR Preprints 28/02/2025:73226)
DOI: https://doi.org/10.2196/preprints.73226

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Effectiveness of Large Language Models in Stroke Rehabilitation Health Education: A Comparative Study of ChatGPT-4, MedGo, Qwen, and ERNIE Bot

Abstract

Background: Stroke is a leading cause of disability and death worldwide, with home-based rehabilitation playing a crucial role in improving patient prognosis and quality of life. Traditional health education models often fall short in terms of precision, personalization, and accessibility. In contrast, large language models (LLMs) are gaining attention for their potential in medical health education, owing to their advanced natural language processing capabilities. However, the effectiveness of LLMs in home-based stroke rehabilitation remains uncertain.

Objective: This study evaluates the effectiveness of four LLMs—ChatGPT-4, MedGo, Qwen, and ERNIE Bot—in home-based stroke rehabilitation. The aim is to offer stroke patients more precise and secure health education pathways while exploring the feasibility of using LLMs to guide health education.

Methods: In the first phase of this study, a literature review and expert interviews identified 15 common questions and 2 clinical cases relevant to stroke patients in home-based rehabilitation. These were input into four LLMs for simulated consultations. Six medical experts (2 clinicians, 2 nursing specialists, and 2 rehabilitation therapists) evaluated the LLM-generated responses using a Likert 5-point scale, assessing accuracy, completeness, readability, safety, and humanity. In the second phase, the top two performing models from phase one were selected. Thirty stroke patients undergoing home-based rehabilitation were recruited. Each patient asked both models 3 questions, rated the responses using a satisfaction scale, and assessed readability, text length, and recommended reading age using a Chinese readability analysis tool. Data were analyzed using one-way ANOVA, post hoc Tukey HSD tests, and paired t-tests.

Results: The results revealed significant differences across the four models in five dimensions: accuracy (*P* = .002), completeness (*P* < .001), readability (*P* = .04), safety (*P* = .007), and humanity (*P* < .001). ChatGPT-4 outperformed all models in each dimension, with scores for accuracy (M = 4.28, SD = 0.84), completeness (M = 4.35, SD = 0.75), readability (M = 4.28, SD = 0.85), safety (M = 4.38, SD = 0.81), and user-friendliness (M = 4.65, SD = 0.66). MedGo excelled in accuracy (M = 4.06, SD = 0.78) and completeness (M = 4.06, SD = 0.74). Qwen and ERNIE Bot scored significantly lower across all five dimensions compared to ChatGPT-4 and MedGo. ChatGPT-4 generated the longest responses (M = 1338.35, SD = 236.03) and had the highest readability score (M = 12.88). In the second phase, ChatGPT-4 performed the best overall, while MedGo provided the clearest responses.

Conclusions: LLMs have shown strong performance in home-based stroke rehabilitation education, demonstrating significant potential for real-world applications. However, further improvements are needed in accuracy, professionalism, and oversight.

Keywords: Large language models; Stroke; Artificial intelligence; Home rehabilitation

Introduction

Stroke is a leading cause of mortality and disability among middle-aged and elderly individuals worldwide. According to the World Health Organization (WHO), over 15 million people experience a stroke annually, resulting in approximately 6 million deaths and leaving millions more with varying degrees of long-term disability [1]. With the global aging population, stroke prevalence continues to rise, particularly in low- and middle-income countries, where the disease burden is increasing, posing a significant public health challenge [2]. Stroke rehabilitation is critical for improving patient prognosis and quality of life [1]. Home-based rehabilitation, offering convenience and cost-effectiveness, has gained significant attention in recent years. However, successful home rehabilitation requires active patient engagement, effective family support, and professional guidance [3]. Comprehensive health education and rehabilitation instructions are essential to ensure patients adhere to appropriate recovery protocols at home. Traditional health education methods, such as printed materials, books, and verbal instructions, often face challenges like delayed information dissemination, inconsistent interpretation, and a lack of personalized support, which can hinder rehabilitation outcomes.

In recent years, large language models (LLMs) have rapidly advanced, gaining significant attention in the medical and healthcare fields due to their advancements in natural language processing (NLP). LLMs possess powerful language comprehension and generation capabilities, enabling them to deliver personalized, easily understandable health guidance tailored to patients’ needs [4,5]. Existing studies have explored LLM applications in patient education, mental health interventions, and health management [6,7]. However, LLMs vary considerably in terms of medical accuracy, comprehensiveness, safety, and readability.

To address these challenges, Shanghai East Hospital developed MedGo in 2024, a specialized Chinese medical LLM designed to assist healthcare professionals in clinical decision-making and medical consultations. MedGo has demonstrated exceptional capabilities in medical task processing, ranking among the top models in the Chinese Biomedical Language Understanding Evaluation (CBLUE) benchmark and excelling in medical question-answering assessments [8]. Given China’s large patient population and limited medical resources, providing personalized home rehabilitation plans for every patient remains a significant challenge. Compared to traditional health education methods, LLMs offer real-time interaction, personalized recommendations, and continuous 24/7 support through intelligent dialogue systems, making them well-suited to meet the information and guidance needs of stroke patients during home rehabilitation.

The application of LLMs in the medical field is still in its early stages [9,10], with limited exploration of how to effectively integrate them into home rehabilitation for stroke patients. Most existing studies [11] primarily focus on general-purpose LLMs (e.g., ChatGPT, Google Bard), overlooking the potential of specialized medical models in healthcare. Furthermore, comprehensive comparisons of various LLMs in medical applications are scarce.

This study aims to evaluate the effectiveness of multiple LLMs, including specialized medical models, in supporting home rehabilitation for stroke patients. By offering a more precise and safer rehabilitation education pathway, this research seeks to advance the application of LLMs in healthcare, providing a more efficient and scientifically sound solution for stroke patients' home rehabilitation. The findings have significant academic and practical implications for promoting the adoption of LLMs in healthcare.

Methods

Study Design

This study was conducted from January 5 to February 22, 2025. Employing both quantitative and qualitative analyses, the study aimed to evaluate the effectiveness of LLMs in home-based stroke rehabilitation education. By comparing the performance of responses from four different LLMs and incorporating expert ratings and patient feedback, we explored their practical application in the rehabilitation process. The detailed procedure is outlined in Figure 1.

Study Subjects

This study selected four LLMs — ChatGPT, MedGo, Qwen, and ERNIE Bot — based on a comparative analysis of general-purpose and medical-specific LLMs, technological diversity, and practical applicability. ChatGPT is a general-purpose model widely used for dialogue generation and question-answering across various domains [12,13]. MedGo, developed by Shanghai Oriental Hospital, is a Chinese medical-specific LLM optimized for healthcare, demonstrating exceptional capabilities in medical decision-making and health consultation. Qwen is an intelligent conversational system that integrates multi-domain knowledge, offering strong multilingual support and advanced question-answering abilities. ERNIE Bot, developed by Baidu, is a large-scale generative model with robust Chinese language understanding and generation capabilities, applicable across a wide range of professional fields (see Table 1).

Table 1. Study Models

Model	Version	Source	Medical-Specific	Open-Source
Chat-GPT	V4.0	Open AI（USA）	No	No
Med-Go	/	Laboratory of Biomedical Artificial Intelligence (China)	Yes	No
Qwen	Qwen-Max	Alibaba (China)	No	No
ERNIE Bot	V3.5	Baidu (China)	No	No

Phase One

Questionnaire Design

The study aimed to evaluate the effectiveness of LLMs in home-based stroke rehabilitation education. By comparing the performance of responses from four different LLMs and incorporating expert ratings and patient feedback, we explored their practical application in the rehabilitation process. The detailed procedure is outlined in Figure 1.

Common inquiries about home rehabilitation for stroke patients from the past three years were gathered and summarized from online medical platforms, including "DingXiangYuan," "HaoDF," "Baidu Health Ask a Doctor," and "AliHealth." Additionally, international stroke rehabilitation guidelines [14] — such as those from the American Heart Association , the American Stroke Association, and the Chinese Stroke Association — were reviewed to identify key concerns during the postoperative rehabilitation phase. Based on this information, an initial set of questions was selected.

Using this feedback and incorporating the clinical experience of neurorehabilitation nursing experts, a set of 15 targeted questions and two common home rehabilitation case scenarios were compiled. Case 1 focused on improving limb mobility and self-care abilities in daily life, with an emphasis on blood pressure control. Case 2 addressed not only limb and gait training but also speech and swallowing rehabilitation, diabetes management, and shoulder pain relief. The questionnaire (see Table 2) covers three key dimensions:

Table 2. Questions and clinical scenarios

No.	Question
Q1	How to recognize a stroke?
Q2	How to prevent a stroke?
Q3	What is the optimal period for stroke rehabilitation?
Q4	How should stroke patients conduct home rehabilitation training after discharge?
Q5	What special considerations should be taken into account during home rehabilitation for stroke patients?
Q6	How should stroke patients and their families manage diet during rehabilitation?
Q7	How should stroke patients manage medication during home rehabilitation?
Q8	How should stroke patients use assistive devices for rehabilitation at home?
Q9	How can the affected limbs of stroke patients be stimulated to promote recovery?
Q10	What functions can patients regain through rehabilitation training?
Q11	Does earlier rehabilitation lead to better outcomes?
Q12	Can stroke patients fully recover through rehabilitation training?
Q13	If no rehabilitation training is conducted, can patients still recover over time?
Q14	Are individuals with a family history of stroke at higher risk of having a stroke?
Q15	Do patients still need regular hospital check-ups during the home rehabilitation period?
Case1	Mr. Shen, male, 71 years old, suffered a stroke two months ago and has been undergoing pharmacological treatment. The patient still experiences hemiparesis, with no active movement in the left upper limb and poor stability in the left upper and lower limbs, requiring assistance when walking. He also needs assistance with daily activities such as bathing. The patient has had hypertension for two years. How should he proceed with home rehabilitation?
Case2	Mr. Wu, male, 77 years old, suffered a stroke one month ago and has been receiving pharmacological treatment. He still has residual weakness in the left side of his body, with impaired fine motor control in the left hand, an unstable gait when walking, difficulty with left-sided weight-bearing, and inability to stand independently. He also experiences frequent nighttime urination and needs assistance with eating, drinking, and bathing. After physical activity, he often feels weakness in his left lower limb. The patient has had hypertension for 40 years and was diagnosed with type 2 diabetes two months ago, with recent fluctuations in blood glucose levels. How should he undergo home rehabilitation?

Model Testing

To enhance the professionalism and specificity of LLM responses, ensuring they address questions from a targeted perspective and align closely with real-world applications, a standardized prompt was added before each model-generated response. The prompt instructed:

"Assume you are an experienced rehabilitation specialist responsible for assessing the home rehabilitation needs of stroke patients and providing personalized advice based on their specific conditions. Use your professional knowledge to provide detailed answers to the following questions. Note that the inquirer is a patient or caregiver with no medical background. Ensure your responses include clear explanations and reference relevant medical evidence to aid understanding."

Fifteen home rehabilitation questions and two clinical cases related to stroke were input into four different LLMs. Each model received the inputs three times, with each iteration conducted in a new conversation to eliminate prior chat history, allowing for an assessment of response consistency. The responses from all models were recorded in plain text format [15]. A single-blind method was applied: the three responses from each model were randomized and grouped into four sets of questions, which were then compiled into a questionnaire for expert evaluation.

Expert Evaluation

A Likert 5-point rating scale was used to assess the outputs of the four LLMs across five dimensions: accuracy, completeness, readability, safety, and humanity

1. Accuracy: Evaluates whether the model-generated information aligns with scientific knowledge and medical facts.
2. Completeness: Assesses whether the model's responses fully cover all relevant aspects of the question or case scenario.
3. Readability: Determines whether the model's language is clear, concise, and easy to understand.
4. Safety: Examines whether the model provides safe and appropriate recommendations without potential risks.
5. Humanity: Assesses whether the model's responses consider patients' emotional needs, dignity, and individual differences while offering care and support.

Note: Experts could provide brief explanations in the comment section if they had concerns about any response. The detailed definitions of the Likert 5-point rating scale are provided in Table 3.

Table 3. Likert 5-point rating scale

Score	Accuracy	Completeness	Readability	Safety	Humanity
1	The response is completely inaccurate, containing misleading or erroneous information.	The response is extremely incomplete, missing key content.	The response is difficult to understand, with unclear and confusing language.	The response contains unsafe health recommendations or content that could mislead patients.	The response is very indifferent, ignoring the emotional needs of patients.
2	The response is partially accurate but contains some errors or incomplete information.	The response is somewhat comprehensive but omits critical or important details.	The response is somewhat understandable but includes ambiguous or unclear parts.	The response is partially safe, but some suggestions may pose potential risks.	The response is somewhat indifferent, lacking empathy for patients.
3	The response is mostly accurate but contains	The response is mostly comprehensive, though	The response is mostly understandable, but some	The response is largely safe, though certain suggestions	The response is mostly human-

	minor errors.	some information is missing.	parts may require further clarification.	may require further verification or revision.	centered but still has room for improvement.
4	The response is generally accurate, with few errors.	The response is generally comprehensive, covering most relevant information.	The response is generally clear and easy to understand.	The response is generally safe and aligns with the patient's health needs.	The response demonstrates empathy and consideration for the patient's emotions.
5	The response is completely accurate, with no errors.	The response is highly comprehensive, covering all relevant information.	The response is highly clear, concise, and well-structured.	The response is entirely safe, with no potential risks.	The response is highly human-centered, full of care and empathy.

A total of six stroke rehabilitation experts (Table 4) were selected from tertiary general hospitals and specialized rehabilitation centers in Shanghai. The expert panel included one stroke center medical specialist, three neurorehabilitation nursing specialists, and two rehabilitation therapists, grouped into three teams [16]. Following a blind evaluation protocol, the experts assessed the outputs of the four LLMs across five dimensions. The evaluation was conducted in three rounds to minimize bias. Prior to the formal assessment, a structured evaluation protocol was developed to guide reviewers through the process and scoring criteria. Clear explanations and definitions for each dimension, along with examples, were provided to ensure understanding. Additionally, the six experts conducted an interpretability analysis of the responses generated by the four LLMs.

Table 4. Expert Profiles

Field of Expertise	Expert	Degree	Title
Clinical Medicine	Expert 1	Ph.D.	Chief Physician
Nursing	Expert 2	Ph.D.	Chief Nurse
Nursing	Expert 3	Bachelor's	Associate Chief Nurse
Nursing	Expert 4	Bachelor's	Associate Chief Nurse
Rehabilitation Medicine	Expert 5	Master's	Senior Therapist
Rehabilitation Medicine	Expert 6	Master's	Senior Therapist

Phase Two

Thirty stroke patients were recruited in Shanghai. The two top-performing models from Phase One were selected for interaction with the patients in a real clinical setting. Each patient asked both models three questions related to home-based stroke rehabilitation. The researchers recorded the responses and independently rated them using a satisfaction scale.

Inclusion Criteria:

1. Patients diagnosed with stroke (including ischemic or hemorrhagic stroke) who meet the diagnostic criteria in the *2021 China Stroke Prevention and Treatment Guidelines* and have been confirmed by CT or MRI imaging.
2. Patients in the rehabilitation phase (1–12 months post-discharge, with ongoing rehabilitation needs).

3. Patients undergoing home-based rehabilitation (not long-term hospitalized).
4. Patients with basic communication skills, able to express their needs accurately (or with a family member to assist in communication).

Exclusion Criteria:

1. Patients with severe cognitive impairments or those unable to accurately express their needs.
2. Patients unable to undergo home-based rehabilitation (e.g., those requiring long-term hospitalization due to the severity of their condition).
3. Patients with other serious comorbidities (e.g., end-stage cancer, severe heart failure) that would interfere with the rehabilitation plan.

The assessment criterion was patient satisfaction, based on the following specific criteria:

Table 5: Patient Satisfaction Rating Scale

Please select the option that best reflects your overall feeling toward the model's response.	
1 Point: Very Dissatisfied	The response from the model is very unsatisfactory, completely failing to meet my needs.
2 Points: Dissatisfied	The response from the model is unsatisfactory, with many issues, and does not meet my needs.
3 Points: Neutral	The response from the model is acceptable; it answers some questions but still has room for improvement.
4 Points: Satisfied	The response from the model is satisfactory, and most of the questions have been answered effectively.
5 Points: Very Satisfied	The response from the model is excellent, and all questions have been answered very well.

Statistical Analysis

Primary Outcomes

Data from the six reviewers' ratings across each dimension for the four sets of responses were compiled and analyzed. Given the small sample size, median values and means ± standard deviations were used as statistical indicators.

In Phase One, expert evaluations of the four LLMs' responses were conducted using a Likert 5-point scale for multi-dimensional assessment. In Phase Two, patient evaluations were performed using a satisfaction scale. The rating data were analyzed using the following statistical methods:

Normality Test: The Shapiro-Wilk test was applied to assess the normality of the rating data distribution, guiding the selection of appropriate methods for subsequent difference testing.

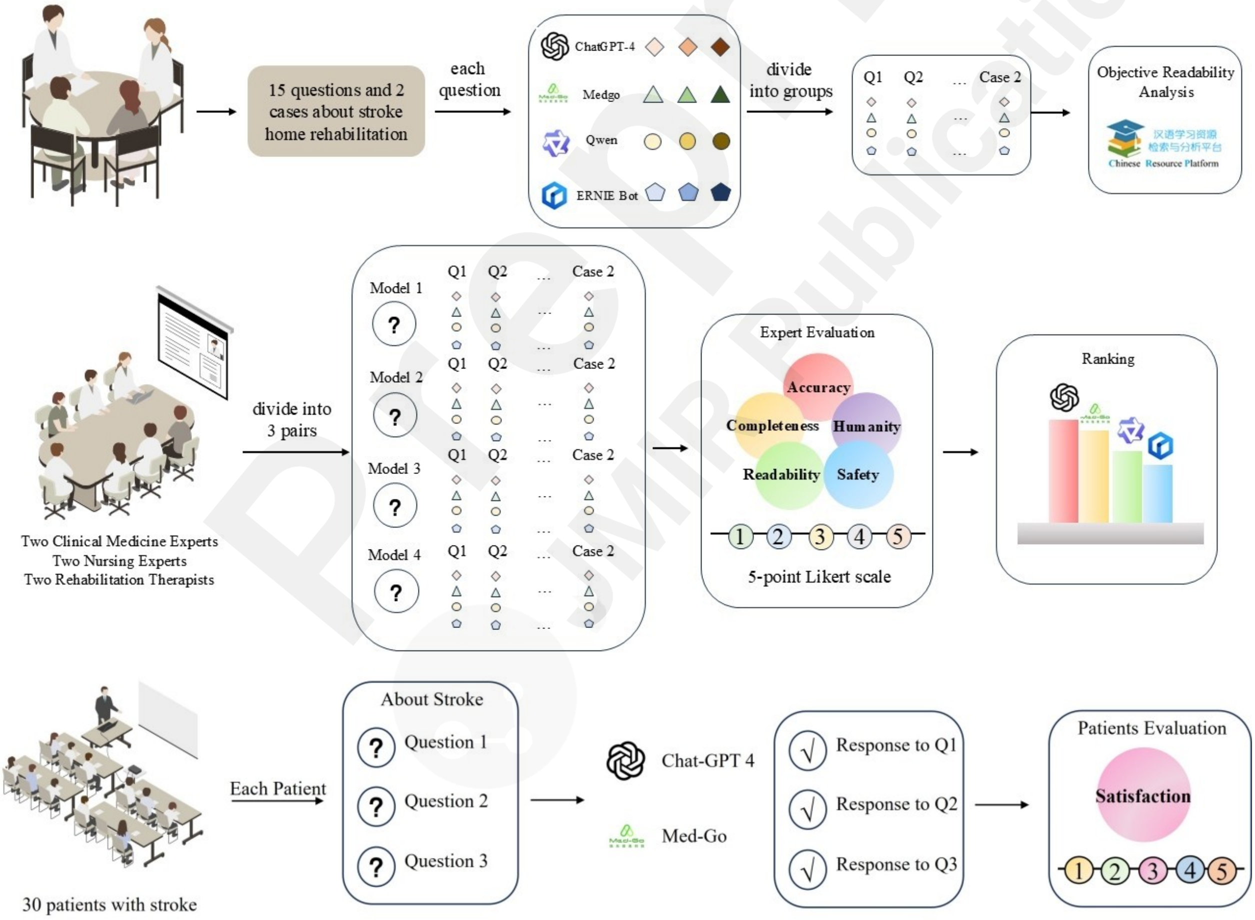
Difference Analysis: For normally distributed data, one-way ANOVA was used in Phase One to compare rating differences between models, followed by post-hoc Tukey HSD analysis. For non-normally distributed data, the Kruskal-Wallis H test and Mann-Whitney U test (with Bonferroni correction) were applied. In Phase Two, paired sample t-tests were used to compare differences between models.

Secondary Outcomes

An objective readability analysis was conducted on the responses generated by the four LLMs using a Chinese Readability Assessment Platform. This online tool [17,18] evaluates text readability by analyzing 52 linguistic features through a multiple linear regression model. The platform provides metrics, including education level, reading difficulty, and recommended reading age, with higher scores indicating more complex text.

A one-way ANOVA was used to assess differences among the four LLMs in terms of word count, reading difficulty scores, and recommended reading age. Post-hoc analysis was performed using Tukey’s HSD test to examine inter-model differences. Additionally, dot plots were generated using the HIPLOT online tool [19] to visually present the readability scores for each model. All statistical analyses were performed using SPSS and HIPLOT, with a significance level set at α = 0.05.

Figure 1. Research Design Workflow Diagram. *Phase One:* Common questions from stroke patients undergoing home-based rehabilitation were collected, and based on the International Stroke Rehabilitation Guidelines and expert input, a questionnaire was developed containing 15 questions and 2 typical cases. These questions and cases were input into four LLMs—ChatGPT, MedGo, Qwen, and ERNIE Bot—with each model receiving the inputs three times. The models' raw text responses were recorded. Two clinical medicine experts, two nursing specialists, and two rehabilitation therapists evaluated the responses using a Likert 5-point scale across five dimensions: accuracy, completeness, readability, safety, and humanity. This evaluation was conducted in three rounds, followed by statistical analysis of the ratings. *Phase Two:* Based on Phase One results, the top two performing LLMs were selected for interaction with 30 patients, who provided satisfaction ratings. [20]



Results

Phase One: Primary Outcomes

Six experts evaluated the responses generated by the LLMs across five dimensions: accuracy, comprehensiveness, readability, safety, and user-centeredness. Figure 2 presents the scores for each LLM.

Among the models, ChatGPT-4 achieved the highest scores across all dimensions, with particularly outstanding performance in safety (M = 4.38, SD = 0.81) and humanity (M = 4.65, SD = 0.65). MedGo performed well in accuracy (M = 4.06, SD = 0.78) and completeness (M = 4.06, SD = 0.74) but was slightly inferior to ChatGPT-4 in the humanity dimension. Qwen and ERNIE Bot received significantly lower scores than both ChatGPT-4 and MedGo.

A one-way ANOVA (see Table 6) revealed significant differences across all five evaluation dimensions: Accuracy: ChatGPT-4 achieved the highest mean score (4.28 ± 0.84), followed by MedGo (4.06 ± 0.78), while Qwen and ERNIE Bot both scored lower (3.91 ± 0.77 and 3.91 ± 0.81, respectively). Completeness: ChatGPT-4 again led (4.35 ± 0.75), followed by MedGo (4.06 ± 0.74), with Qwen and ERNIE Bot receiving lower scores. Readability: ChatGPT-4 scored the highest (4.28 ± 0.85), followed by MedGo (4.17 ± 0.81) and Qwen (4.02 ± 0.81), while ERNIE Bot had the lowest score (3.99 ± 0.79). Safety: ChatGPT-4 topped this dimension (4.38 ± 0.81), followed by MedGo (4.23 ± 0.73), Qwen (4.08 ± 0.78), and ERNIE Bot (4.05 ± 0.75). Humanity: ChatGPT-4 achieved the highest score (4.65 ± 0.66), followed by MedGo (4.38 ± 0.73), with Qwen and ERNIE Bot both scoring identically (4.27 ± 0.72 and 4.27 ± 0.75, respectively).

Figure 4 illustrates the performance of different LLMs across the five evaluation dimensions. ChatGPT-4 exhibited the best overall performance in all dimensions. The line chart shows the variation in average scores, with each line representing the scoring trend of a model across different questions. The radar chart offers a visual representation of each model's performance across the five dimensions. Each axis of the radar chart corresponds to an evaluation dimension, with a larger area indicating stronger performance in that dimension.

Phase One: Secondary Outcomes

Descriptive statistics for the objective readability analysis are presented in Table8-9 and Figure 5. Chinese Character Count: ChatGPT-4 generated the highest total character count (22,752 characters) and the highest average word count (1338.35 ± 236.03). In contrast, Qwen produced the shortest text, with a total of 13,481 characters and an average word count of 793.00 ± 283.64. Reading Difficulty Score: ChatGPT-4 had the highest average reading difficulty score (12.88), while ERNIE Bot had the lowest (11.92). Recommended Reading Age: ChatGPT-4 also had the highest mean recommended reading age (12.82), whereas ERNIE Bot had the lowest (11.94).

The results of the one-way ANOVA are shown in Table 6: Chinese Character Count: Significant differences were observed in the total text length among the LLMs (F = 11.43 > F crit = 2.75, *P* < .001), with the most significant difference between ChatGPT-4 and Qwen (*P* < .001). Reading Difficulty Score: A significant difference was found in reading difficulty scores among the models (F = 3.32 > F crit = 2.75, *P* = .03). Recommended Reading Age: No significant differences were observed in recommended reading age among the models (F = 2.48 < F crit = 2.75, *P* = .07).

Phase Two: Patient Interaction Results

A total of 30 eligible patients were recruited, generating 90 questions (Figure 6). These, along with expert-suggested questions, were categorized into seven groups: basic definitions and types, causes and risk factors, daily management and care, rehabilitation training, effectiveness and safety, emotional support, complications, and rehabilitation environment and equipment.

Paired sample t-tests showed that the average score for GPT was slightly higher than that for MedGo (Table 7), with a significant difference between the two models (*P* = .01).

Table 6. One-Way ANOVA of Objective Readability

Dimension	F value	*P* value	F crit
Accuracy	4.93	=.002	2.63
Completeness	7.01	□.001	2.63
Readability	2.87	=.04	2.63
Safety	4.06	=.007	2.63
Humanity	6.18	□.001	2.63
Chinese character count	11.43	□.001	2.75
Reading difficulty score	3.32	=.03	2.75
Recommanded reading age	2.48	=.07	2.75

Table 7. Descriptive Statistics of Patient Interactions

Model	Mean ± SD	T value	*P* value	95% CI
ChatGPT-4	3.34 ± 0.64	2.65	.01	[0.08, 0.53]
MedGo	3.04 ± 0.78			

Table 8. Descriptive Statistics of Expert Rating Analysis and Objective Readability Analysis (1)

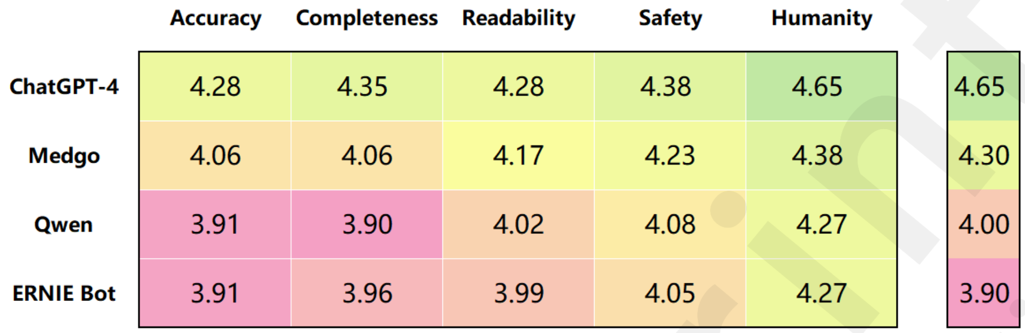
Dimension	Chat-GPT4		MedGo	
	Median(IQR)	Mean±SD	Median(IQR)	Mean±SD
Accuracy	5.0(2.0-5.0)	4.28 ± 0.84	4.0(2.0-5.0)	4.06 ± 0.78
Completeness	5.0(3.0-5.0)	4.35 ± 0.75	4.0(2.0-5.0)	4.06 ± 0.74
Readability	5.0(3.0-5.0)	4.28 ± 0.85	4.0(2.0-5.0)	4.17 ± 0.81
Safety	5.0(2.0-5.0)	4.38 ± 0.81	4.0(3.0-5.0)	4.23 ± 0.73
Humanity	5.0(3.0-5.0)	4.65 ± 0.65	5.0(3.0-5.0)	4.27 ± 0.73
Chinese character count	1367.00(916.00-1697.00)	1338.35 ± 236.03	998.00(726.00-1470.00)	1048.35 ± 195.26
Reading difficulty score	12.81(11.13-15.16)	12.88 ± 0.82	12.30(11.21-14.52)	12.38 ± 0.90
Recommanded reading age	13.00(11.00-15.00)	12.82 ± 0.78	12.00(11.00-14.00)	12.29 ± 0.85

Table 9. Descriptive Statistics of Expert Rating Analysis and Objective Readability Analysis (2)

Dimension	Qwen		ERNIE Bot	
	Median(IQR)	Mean±SD	Median(IQR)	Mean±SD
Accuracy	4.0(2.0-5.0)	3.91 ± 0.77	4.0(2.0-5.0)	3.91 ± 0.81
Completeness	4.0(2.0-5.0)	3.90 ± 0.78	4.0(2.0-5.0)	3.96 ± 0.78
Readability	4.0(2.0-5.0)	4.02 ± 0.81	4.0(2.0-5.0)	3.99 ± 0.79

Safety	4.0(2.0-5.0)	4.08 ± 0.80	4.0(2.0-5.0)	4.05 ± 0.75
Humanity	4.0(2.0-5.0)	4.27 ± 0.72	4.0(3.0-5.0)	4.27 ± 0.75
Chinese character count	772.00(325.00-1223.00)	793.00 ± 283.64	867.00(694.00-2228.00)	979.12 ± 361.84
Reading difficulty score	12.40(11.26-14.01)	12.28 ± 0.71	11.95(10.09-14.15)	11.92 ± 1.09
Recommanded reading age	12.00(11.00-14.00)	12.18 ± 0.78	12.00(10.00-14.00)	11.94 ± 1.43

Figure 2. Heatmap of the Average Scores for Responses from the Four LLMs



	Accuracy	Completeness	Readability	Safety	Humanity	
ChatGPT-4	4.28	4.35	4.28	4.38	4.65	4.65
Medgo	4.06	4.06	4.17	4.23	4.38	4.30
Qwen	3.91	3.90	4.02	4.08	4.27	4.00
ERNIE Bot	3.91	3.96	3.99	4.05	4.27	3.90

Figure 3. Bar Chart of the Average Scores for Responses from the Four LLMs

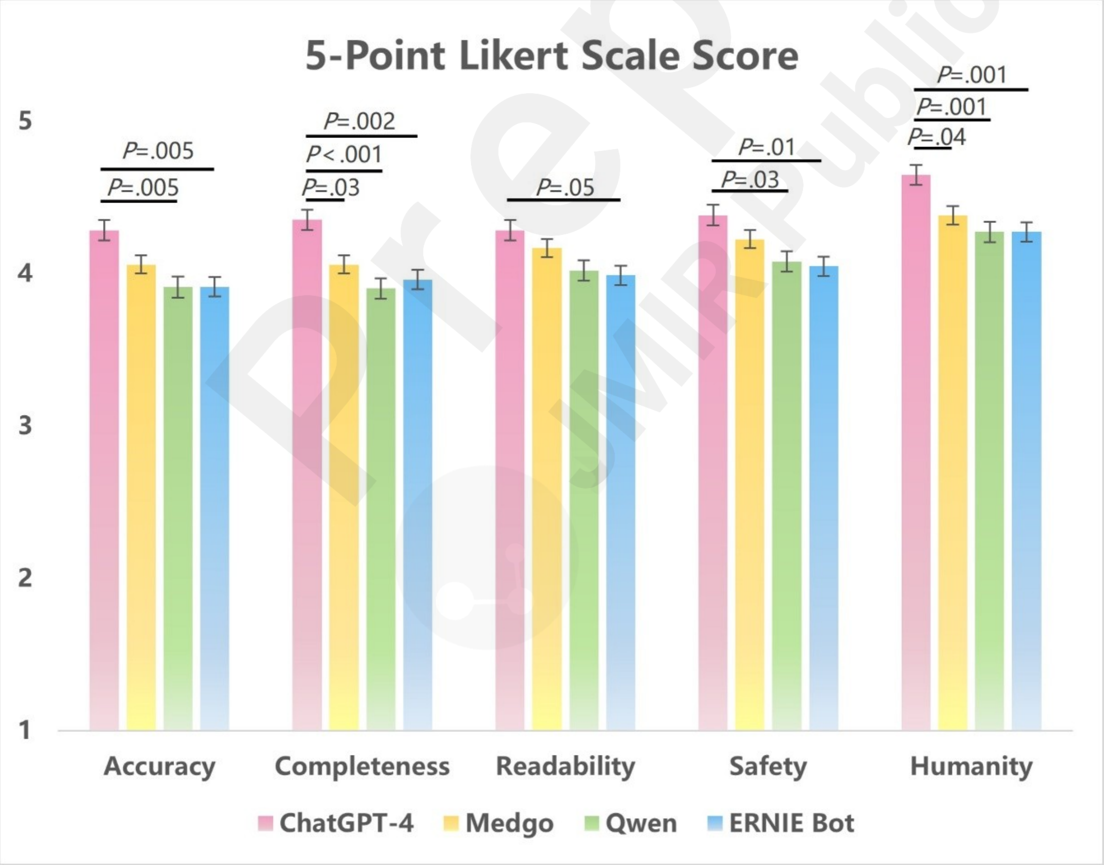


Figure 4. Line Chart of the Median Scores and Radar Chart for the Four LLMs. (A) Accuracy. (B) Completeness. (C) Humanity. (D) Readability. (E) Safety. (F) Radar Chart

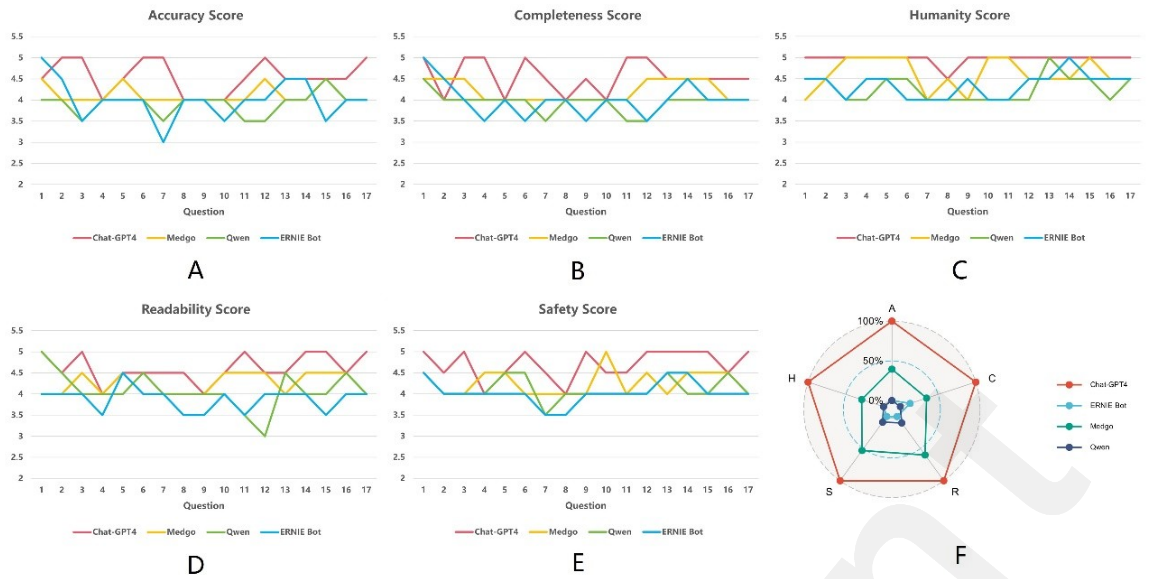


Figure 5. Comparative Evaluation of LLM Responses on Relevant Questions. (A) Box plot showing the variation in text length among the four LLMs, with a significant difference observed between ChatGPT-4 and Qwen (p < .001). (B) Box plot illustrating the variation in reading difficulty scores among the four LLMs. (C) Box plot showing the variation in recommended reading age among the four LLMs (p = .07). (D) Density plot displaying the distribution of reading difficulty scores among the models. (E) Bar chart presenting the distribution of educational levels required to comprehend the responses. All rating data in this study were tested and found to follow a normal or approximately normal distribution.

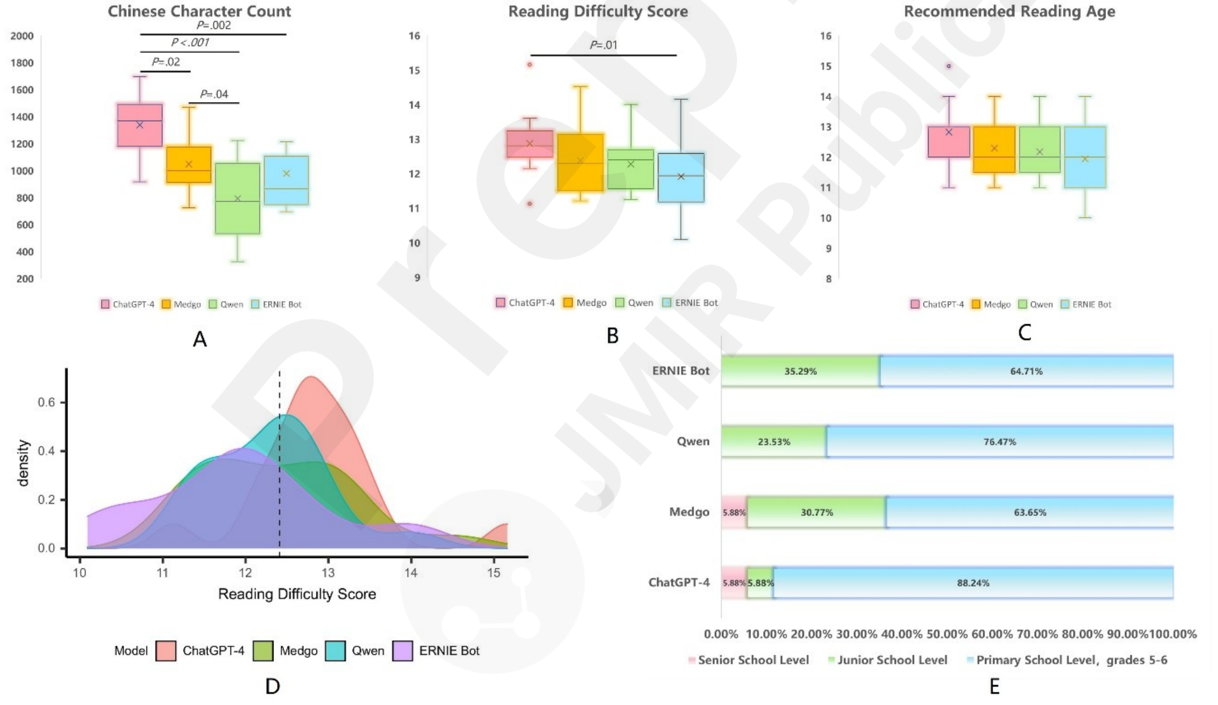
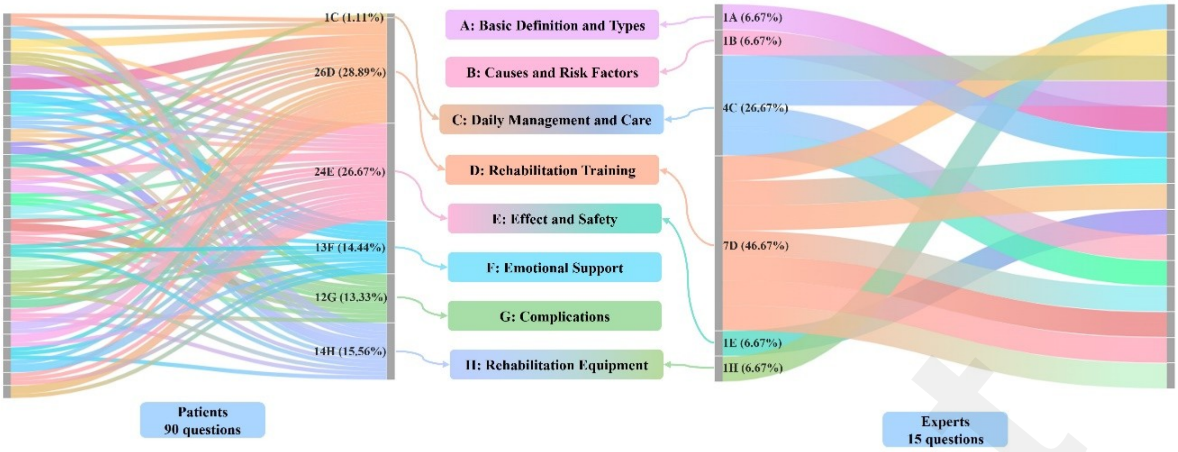


Figure 6: The Sankey diagram illustrates the classification of questions in both phases. On the left, 90 questions posed by 30 patients are shown, while on the right, the 15 integrated questions are displayed.



Discussion

Principal Results

This study evaluated the performance of ChatGPT-4, MedGo, Qwen, and ERNIE Bot in providing health education for stroke patients undergoing home rehabilitation. The results are as follows:

In Phase One, ChatGPT-4 demonstrated the best overall performance across all dimensions, excelling particularly in humanity and safety. This finding aligns with previous studies [21]. MedGo, as a medical-specific model, excelled in accuracy and completeness, underscoring its potential for medical text processing and generation. Qwen and ERNIE Bot received lower scores across all five dimensions compared to ChatGPT-4 and MedGo, indicating a significant performance gap. ChatGPT-4 showed a high and concentrated reading difficulty score distribution, making it well-suited for scenarios that require complex content generation. However, this may present readability challenges for general users. ERNIE Bot and MedGo exhibited lower and more stable readability scores, suggesting they produce easier-to-read content, making them more suitable for general users or tasks that demand lower reading difficulty. Qwen displayed a wide range of readability scores, reflecting greater variability in reading difficulty, but with relatively lower stability compared to the other models.

In Phase Two of patient interactions, ChatGPT-4 received higher ratings than MedGo, but overall, the ratings were lower than those given by the expert group. This discrepancy is mainly due to the challenges patients face in evaluating the models, including variations in personal understanding, needs, and the models' performance and applicability. This is particularly evident when dealing with complex medical information. As AI in medical decision-making is still developing, patients tend to be more skeptical of the models' accuracy and reliability, often finding their responses unclear. In contrast, experts, with their accumulated knowledge and familiarity with medical terminology, are better equipped to interpret the models' medical information, resulting in higher ratings.

There are significant differences in the areas of focus between patients and experts. Patients tend to prioritize rehabilitation methods, outcomes and safety, emotional support, and equipment-related concerns, reflecting their practical needs and psychological state during rehabilitation. Many stroke patients are primarily concerned with improving their quality of life through daily management and care. Given the psychological pressures they face during rehabilitation, emotional support is also crucial. In contrast, experts focus on the effectiveness of rehabilitation plans, safety in technical aspects, and the dissemination of theoretical knowledge. As professionals, they are more likely to base treatment and rehabilitation plans on scientific evidence.

This disparity highlights the communication gap between experts and patients in healthcare. Patients may struggle to fully understand certain medical terms and treatment approaches, leading to confusion or anxiety about their rehabilitation plans. While experts emphasize treatment outcomes and safety, they must also consider how to effectively communicate this specialized knowledge to patients, fostering a correct understanding of rehabilitation and improving treatment adherence.

According to standardized prompts, each LLM was asked to provide sources for their responses. ChatGPT-4 did not explicitly cite references, but its answers, based on its extensive training dataset, generally aligned with medical knowledge and clinical practice. In contrast, MedGo provided more detailed medical support, citing specific medical literature, treatment guidelines, and clinical studies. However, the responses from Qwen and ERNIE Bot lacked clear citations of literature and concrete clinical evidence.

Some responses from the LLMs contained significant errors. For example, in Question 3, which addressed the optimal period for stroke rehabilitation, Qwen incorrectly stated that the chronic phase of stroke begins three months after onset. According to various medical guidelines, the chronic phase typically starts six months after a stroke, making Qwen's response inconsistent with these guidelines. In Question 7, concerning commonly used medications during home-based stroke rehabilitation, ERNIE Bot provided an incorrect answer, mentioning alteplase, a thrombolytic drug that cannot be taken at home and must be administered intravenously in a hospital setting. The use of alteplase requires professional monitoring and must be administered within 4.5 hours of stroke onset.

Analysis revealed that ChatGPT-4 made fewer errors, though it occasionally produced "hallucinations" due to issues with patient language expression. MedGo demonstrated high accuracy but lacked personalized care. Qwen and ERNIE Bot provided incomplete and vague responses.

The errors observed in the LLMs could impact patient rehabilitation to some extent. This is primarily because LLMs generate responses based on statistical language models rather than true understanding of the questions. They lack genuine comprehension and reasoning abilities. Their training depends heavily on large volumes of open-text data from the internet, which does not guarantee the quality or timeliness of the answers. Medical knowledge is vast and complex, and the capabilities of LLMs vary. General-purpose LLMs struggle with specialized medical language and often lack explainability. These models are particularly prone to errors in areas such as disease diagnosis, drug effects, and emerging medical issues.

Therefore, when addressing medical questions, especially in healthcare decision-making, reliance on AI models should be approached with caution. Professional medical judgment remains irreplaceable, particularly when it concerns patient health and treatment plans.

Comparison with Prior Work

Previous studies have shown that LLMs offer benefits in education [22,23], with ChatGPT-4 demonstrating particular advantages in the comprehensibility of medical information [24].

However, most studies focus primarily on the accuracy of medical knowledge responses and applications in medical education [25][26], with limited research on using LLMs for patient health education. This study is the first to comprehensively evaluate the performance of both general-purpose and medical-specific LLMs in home-based stroke rehabilitation health education, providing real-world feedback from patients and family caregivers. MedGo, as a medical-specific LLM, provides clear sources for its answers, which helps reduce the likelihood of errors. However, due to the limitations of the model and the influence of its training data, caution is still necessary, particularly when addressing complex or uncommon medical issues.

The creation and evaluation standards for medical LLMs must be actively developed by the medical community [27] and validated through real-world experiments. While many studies focus on the performance of general-purpose LLMs [28,29], research on medical-specific models remains limited. For instance, some studies have found that Med-PaLM 2 has made significant progress in medical question answering, particularly across multiple medical benchmarks and real-world problem-solving [30]. However, this model is not tailored for the Chinese medical context. Therefore, our study emphasizes the unique value of China-specific medical models, such as MedGo, in improving the quality of healthcare education and providing more professional and personalized solutions, thus addressing a gap in the existing literature.

Limitations

Although this study provides valuable insights into the application of LLMs in home-based stroke rehabilitation health education, several limitations should be noted. First, the expert sample size was small, with only six experts participating in the ratings, which may have affected the comprehensiveness and representativeness of the evaluations. Second, to facilitate patient understanding and streamline the rating process, only a satisfaction scale was used in the second phase of interaction, resulting in a simplified rating criterion. Third, the question design did not address issues related to patient emotions and adherence, despite recognizing during the analysis that emotional changes and patient compliance are crucial factors in home rehabilitation. Fourth, the broader implementation of LLMs requires careful consideration of their economic viability, feasibility, and sustainability. While this study focused primarily on the academic perspective, real-world applications must address cost-effectiveness, technological barriers, and the potential impact on healthcare institutions. Finally, the health education content generated by LLMs still presents potential biases, inconsistent information, and a lack of explainability. Research [31] indicates that the use of LLMs in healthcare faces challenges related to information reliability, biases, ethical compliance, and patient acceptance. Further optimization of model algorithms is needed to improve the reliability of medical knowledge bases, and stronger oversight and regulation of AI-generated health information are essential [32].

Future research on the application of LLMs in healthcare should consider increasing the number of experts or using multi-center data to enhance the reliability of evaluation results. A unified rating standard should be established when comparing expert and patient ratings to ensure result comparability. To address potential resource challenges in LLM application, feasibility assessments should be conducted regarding their management, use, and long-term maintenance. Specifically, for personalized home rehabilitation education for stroke patients, future studies could expand the sample size and include evaluations from diverse patient groups, further exploring the adaptability of LLMs at different stages of rehabilitation. While this study focused on text-based LLMs in stroke rehabilitation education, future developments in multimodal LLMs may enhance patient interaction by incorporating image and speech recognition capabilities [33]. multimodal LLMs could integrate various data sources, potentially improving assessment accuracy and patient care. These efforts will help optimize LLM performance in medical settings and strengthen their role in patient-centered rehabilitation education.

Conclusions

This study evaluated the application of ChatGPT-4, MedGo, ERNIE Bot, and Qwen in home-based stroke rehabilitation health education through a two-stage experiment. The results showed that ChatGPT-4 performed the best, effectively addressing patients' emotional needs and providing personalized recommendations. MedGo, trained on medical-specific data, provided clear and reliable sources for its responses, while the other two general-purpose models performed moderately. A noticeable gap was observed between patient and expert evaluations, underscoring the need for improvements in the accuracy and professionalism of LLMs. Future research should focus on combining the user-friendly aspects of general-purpose models with the accuracy of medical-specific models to enhance the development of intelligent, personalized, and efficient rehabilitation education.

Acknowledgements

This study was funded by three projects: the "2022 Pudong New District Health System Leading Talent Cultivation Program (4-year), grant number PWRd2022-16," the "Tongji University School of Nursing (in preparation) Discipline Construction Three-Year Action Plan (2022-2025), Leading Talent Project, grant number JS2210204," and the "Pudong New District Health Commission, Revision of the Quick Aphasia Assessment Tool (QAB) and its Application in Post-Stroke Aphasia Patients, grant number PW2022A-0."

During the translation process of this manuscript, ChatGPT-4 was used to assist in polishing the text and checking for errors. Apart from this, all other aspects of the study were completed independently by the authors.

The funding for this study was provided by Tongji University and Shanghai East Hospital (project fund of 100,000 RMB). We would like to thank the individuals and organizations mentioned above for their support of this research.

Some icons in picture 1 were sourced from Bioicons [20], and all icons provided by the website are licensed for free use under an open-access license (CC-BY 4.0 or other applicable licenses), with all usage adhering to the relevant licensing requirements.

Authors' Contributions

Conceptualization: Qiang shiqi (lead), Liao yang (equal)
Data curation: Qiang shiqi
Formal analysis: Qiang shiqi (lead), Gu yanfen (supporting) , Zhang yue(supporting)
Funding acquisition: Yu haipin
Investigation: Liao yang(lead), Wang yiyan (equal), Hannuo(supporting)
Methodology: Qiang shiqi, Yu haipin (supporting), Zhang haitao (supporting)
Resources: Liao yang(lead), Wang yiyan (supporting), Zhang yue(supporting)
Validation: Liao yang(lead), XuZehui(supporting)

Visualization: Qiang shiqi
Writing – original draft: Qiang shiqi
Writing – review & editing: Zhang haitao (lead), Yu haipin(equal)

Conflicts of Interest

none declared

Abbreviations

LLMs: Large Language Models

Multimedia Appendix 1

References

1. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, Fisher M, Pandian J, Lindsay P. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke. 2022 Jan;17(1):18-29. doi: 10.1177/17474930211065917. PMID: 34986727
2. Markus HS, Brainin M, Fisher M. Tracking the global burden of stoke and dementia: World Stroke Day 2020. Int J Stroke. 2020 Oct;15(8):817-818. doi: 10.1177/1747493020959186. PMID: 33115386]
3. Bartoli D, Petrizzo A, Vellone E, Alvaro R, Pucciarelli G. Impact of telehealth on stroke survivor-caregiver dyad in at-home rehabilitation: A systematic review. J Adv Nurs. 2024 Oct;80(10):4003-4033. doi: 10.1111/jan.16177. PMID: 38563582
4. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45. doi:10.1145/3641289
5. Denecke K, May R; LLMHealthGroup; Rivera Romero O. Potential of Large Language Models in Health Care: Delphi Study. J Med Internet Res. 2024 May 13;26:e52399. doi: 10.2196/52399. PMID: 38739445
6. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. J Med Internet Res. 2023 Oct 30;25:e49324. doi: 10.2196/49324. PMID: 37902826
7. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content. J Med Internet Res. 2024 Apr 25;26:e55847. doi: 10.2196/55847. PMID: 38663010
8. Zhang H, An B. MedGo: A Chinese Medical Large Language Model[J]. arXiv preprint arXiv:2410.20428, 2024.
9. Lee J H, Choi E, McDougal R, et al. Large Language Model (GPT-4) Accurately Localizes Stroke Lesions (P8-4.002)[C]//Neurology. Hagerstown, MD: Lippincott Williams & Wilkins, 2024, 102(17_supplement_1): 2563. doi:10.1212/wnl.0000000000204601
10. Lehnen NC, Dorn F, Wiest IC, Zimmermann H, Radbruch A, Kather JN, Paech D. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. Radiology. 2024 Apr;311(1):e232741. doi: 10.1148/radiol.232741. PMID: 38625006
11. Neo JRE, Ser JS, Tay SS. Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers. Front Digit Health. 2024 May 9;6:1395501. doi: 10.3389/fdgth.2024.1395501. PMID: 38784703

12. Li KD, Fernandez AM, Schwartz R, Rios N, Carlisle MN, Amend GM, Patel HV, Breyer BN. Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study. J Med Internet Res. 2024 Aug 21;26:e56500. doi: 10.2196/56500. PMID: 39167785
13. Deiner MS, Deiner NA, Hristidis V, McLeod SD, Doan T, Lietman TM, Porco TC. Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study. J Med Internet Res. 2024 Mar 1;26:e49139. doi: 10.2196/49139. PMID: 38427404
14. Klijn CJ, Hankey GJ; American Stroke Association and European Stroke Initiative. Management of acute ischaemic stroke: new guidelines from the American Stroke Association and European Stroke Initiative. Lancet Neurol. 2003 Nov;2(11):698-701. doi: 10.1016/s1474-4422(03)00558-1. PMID: 14572738.
15. Shi R, Liu S, Xu X, Ye Z, Yang J, Le Q, Qiu J, Tian L, Wei A, Shan K, Zhao C, Sun X, Zhou X, Hong J. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. Heliyon. 2024 Jul 14;10(14):e34391. doi: 10.1016/j.heliyon.2024.e34391. PMID: 39113991.
16. Sezgin E, Chekeni F, Lee J, Keim S. Clinical Accuracy of Large Language Models and Google Search Responses to Postpartum Depression Questions: Cross-Sectional Study. J Med Internet Res. 2023 Sep 11;25:e49240. doi: 10.2196/49240. PMID: 37695668.
17. Cheng Y, Xu DK, Dong J. Analysis of key factors in text readability grading and study of readability formula based on Chinese textbook corpus. Appl ling 2020.p.132–43.
18. Chinese Readability Platform. Online text readability analysis tool [Internet].URL: http://120.27.70.114:8000/analysis_a [accessed 2025-01-15]
19. HIPLOT. Online tool for data visualization and statistical analysis [Internet]. URL: https://hiplot.org [accessed 2025-01-20].
20. Bioicons. Free open-source icons for scientific use [Internet]. Bioicons.com; URL: https://bioicons.com/ [accessed 2025-01-10]
21. Cho S, Lee M, Yu J, Yoon J, Choi JB, Jung KH, Cho J. Leveraging Large Language Models for Improved Understanding of Communications With Patients With Cancer in a Call Center Setting: Proof-of-Concept Study. J Med Internet Res. 2024 Dec 11;26:e63892. doi: 10.2196/63892. PMID: 39661975
22. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, Bin Saleh K, Alowais SA, Alshaya OA, Rahman I, Al Yami MS, Albekairy AM. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm. 2023 Aug;19(8):1236-1242. doi: 10.1016/j.sapharm.2023.05.016. PMID: 37321925.
23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198. PMID: 36812645
24. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large Language Models May Help Patients Understand Peer-Reviewed Scientific Articles About Ophthalmology: Development and Usability Study. J Med Internet Res. 2024 Dec 24;26:e59843. doi: 10.2196/59843. PMID: 39719077
25. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS, Hong HJ, Kashyap M, Chaurasia AR, Shah NR, Singh K, Tazbaz T, Milstein A, Pfeffer MA, Shah NH. Testing and Evaluation of Health Care

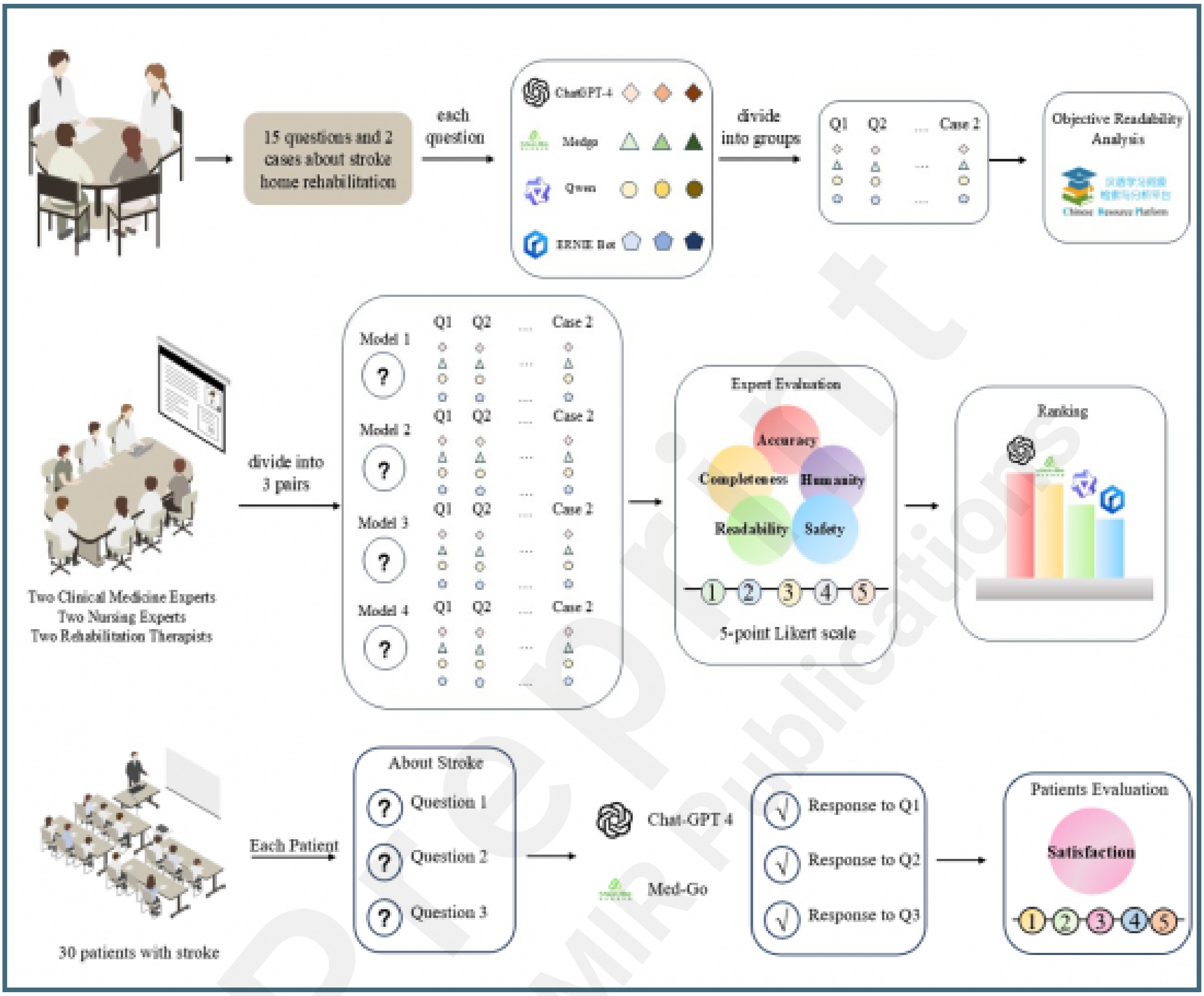
Applications of Large Language Models: A Systematic Review. JAMA. 2025 Jan 28;333(4):319-328. doi: 10.1001/jama.2024.21700. PMID: 39405325

26. Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274. doi: 10.1016/j.lindif.2023.102274
27. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023 Sep 5;330(9):866-869. doi: 10.1001/jama.2023.14217. PMID: 37548965.
28. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne). 2024 Jun 20;11:1380148. doi: 10.3389/fmed.2024.1380148. PMID: 38966538
29. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res. 2023 May;104(5):269-273. doi: 10.4174/astr.2023.104.5.269. PMID: 37179699
30. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025 Jan 8. doi: 10.1038/s41591-024-03423-7. Epub ahead of print. PMID: 39779926.
31. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, Malin B, Yin Z. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. J Med Internet Res. 2024 Nov 7;26:e22769. doi: 10.2196/22769. PMID: 39509695
32. Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. JAMA. 2023 Jul 25;330(4):315-316. doi: 10.1001/jama.2023.9651. Erratum in: JAMA. 2023 Sep 12;330(10):974. doi: 10.1001/jama.2023.16286. PMID: 37410482.
33. Meskó B. The Impact of Multimodal Large Language Models on Health Care's Future. J Med Internet Res. 2023 Nov 2;25:e52865. doi: 10.2196/52865. PMID: 37917126

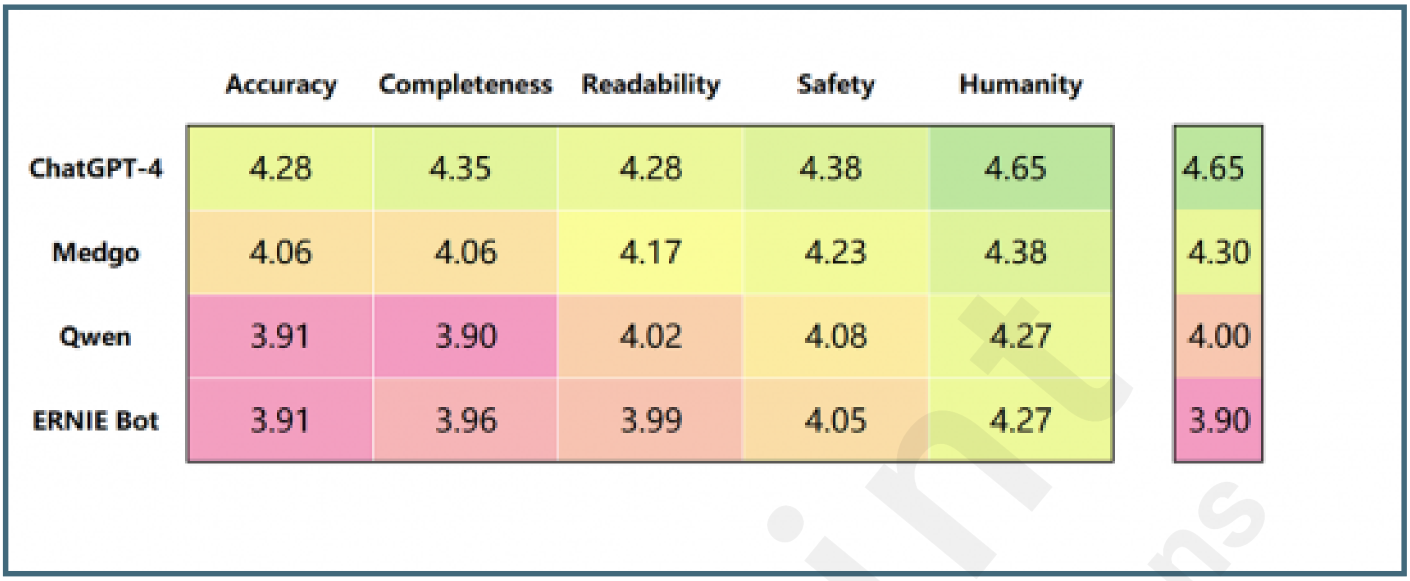
Supplementary Files

Figures

Research Design Workflow Diagram.

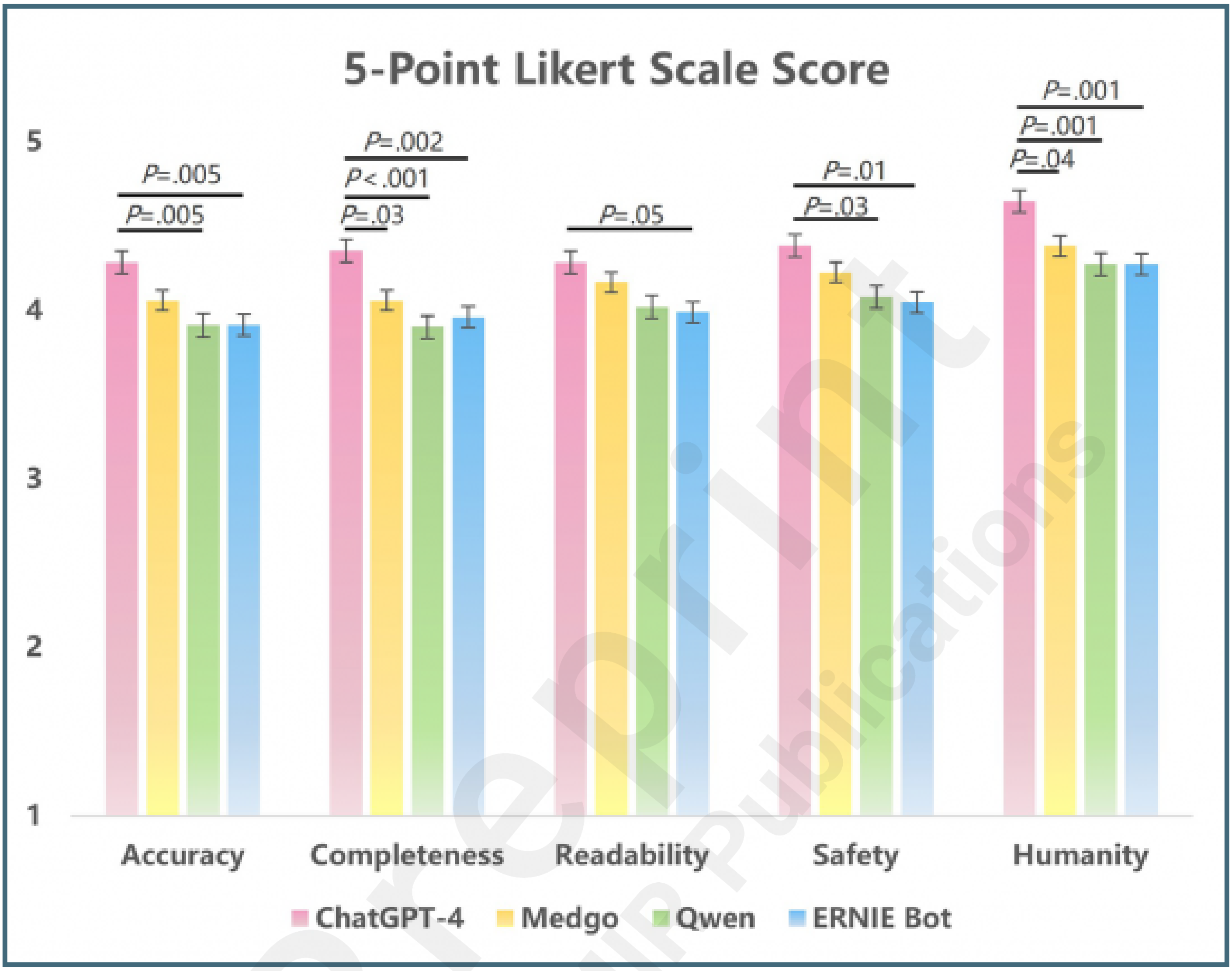


Heatmap of the Average Scores for Responses from the Four LLMs.

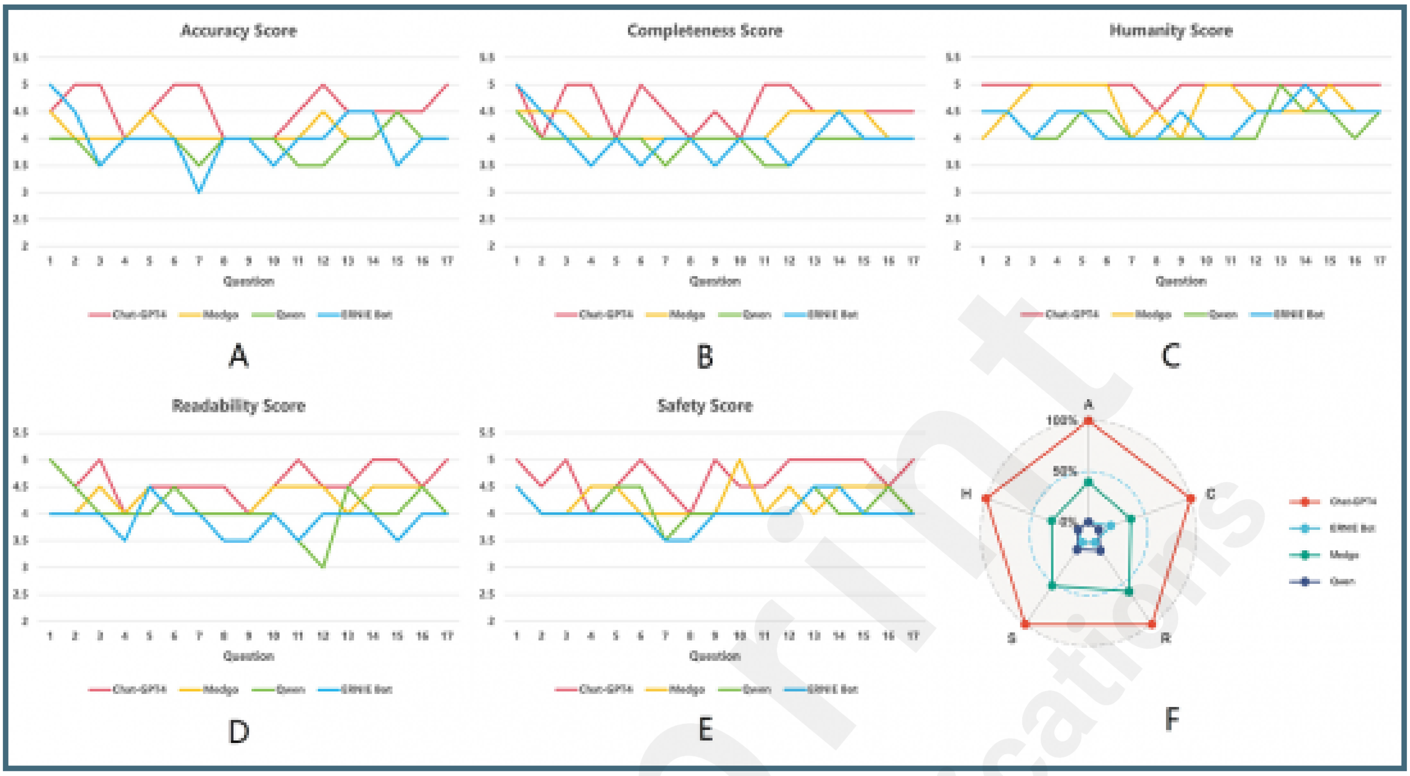


	Accuracy	Completeness	Readability	Safety	Humanity
ChatGPT-4	4.28	4.35	4.28	4.38	4.65
Medgo	4.06	4.06	4.17	4.23	4.38
Qwen	3.91	3.90	4.02	4.08	4.27
ERNIE Bot	3.91	3.96	3.99	4.05	4.27

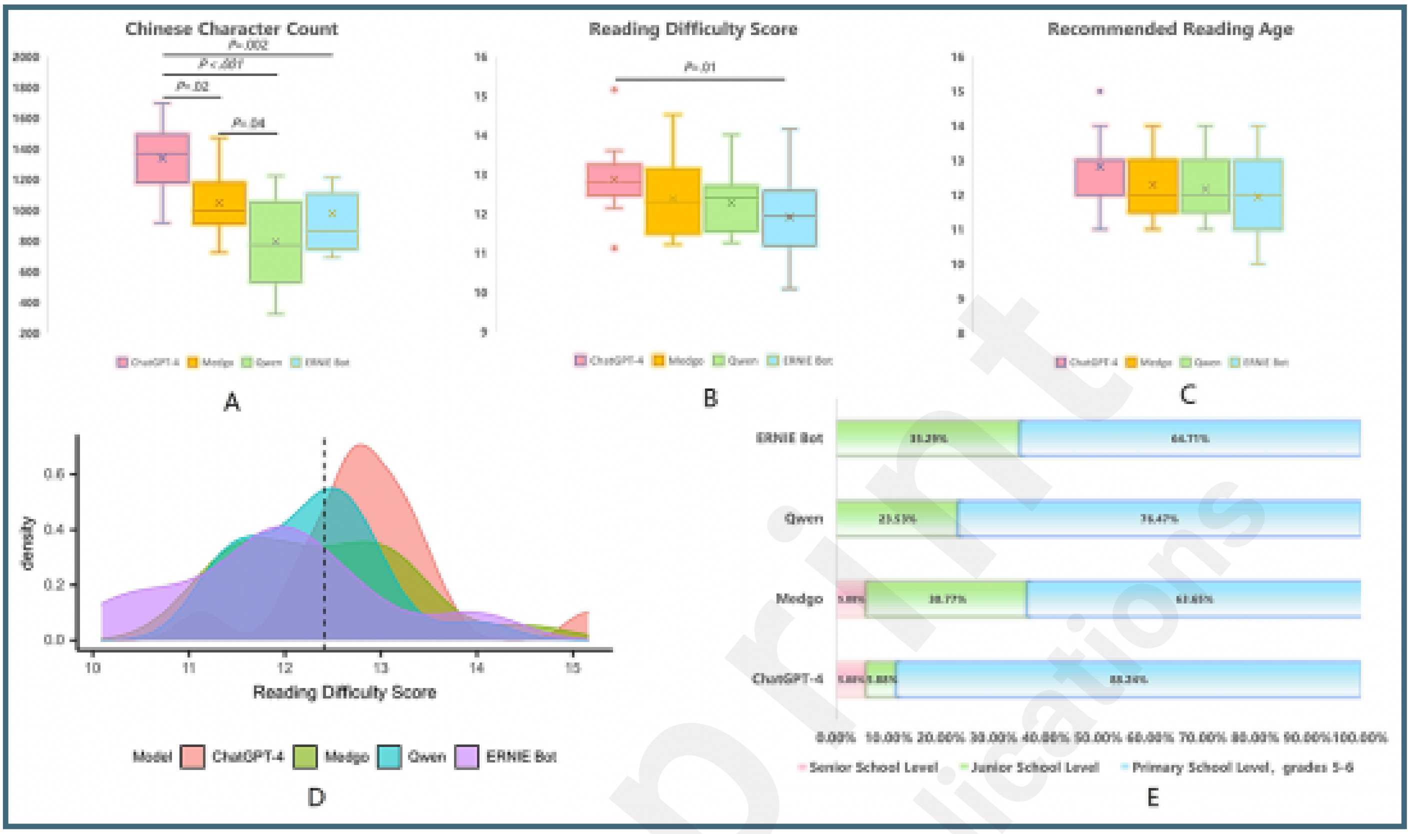
Bar Chart of the Average Scores for Responses from the Four LLMs.



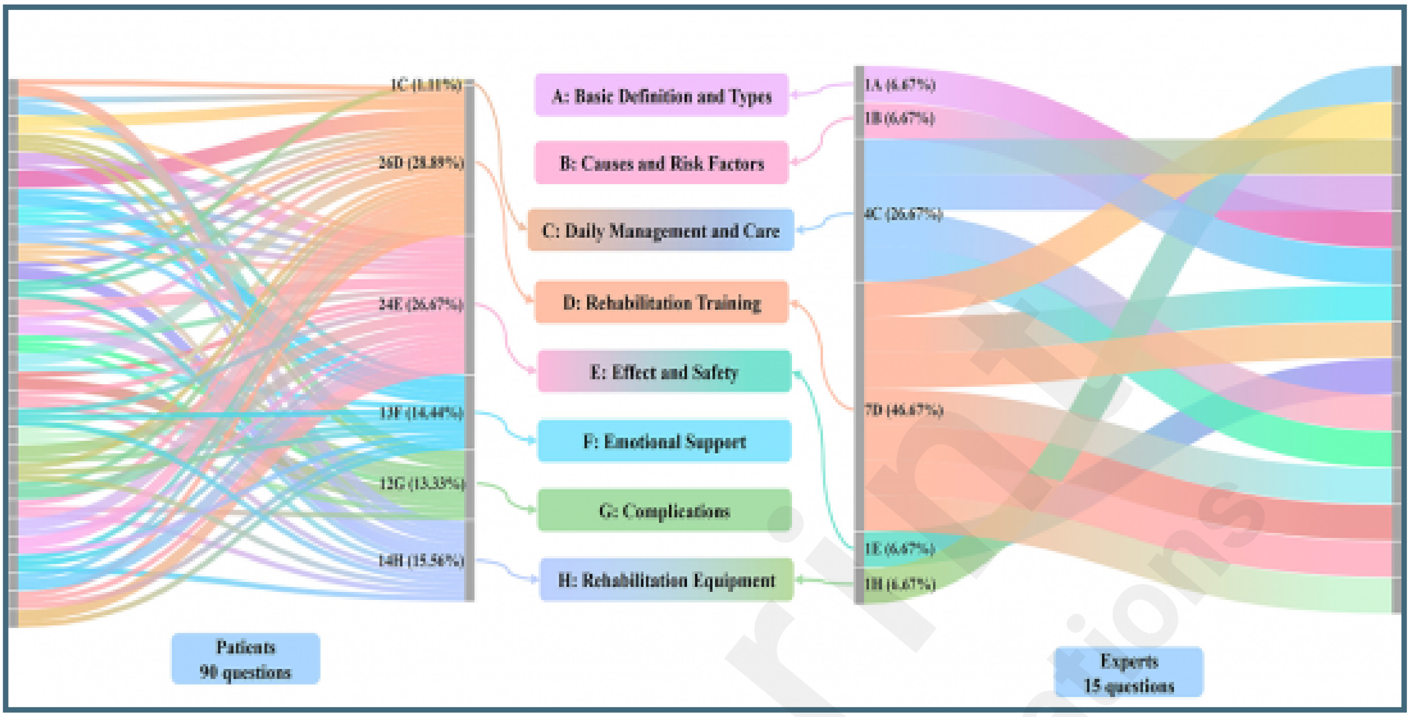
Line Chart of the Median Scores and Radar Chart for the Four LLMs.



Comparative Evaluation of LLM Responses on Relevant Questions.



The Sankey diagram illustrates the classification of questions in both phases.



Multimedia Appendixes

Phase One Expert Scoring Data.
URL: http://asset.jmir.pub/assets/4f1371ab3796eef29e60dfc32883bfa6.xlsx

Phase Two Patient Questions and Scoring Data.
URL: http://asset.jmir.pub/assets/24a60d84804e2061e35710aa552a40ff.xlsx

Ethics Committee Approval Document.
URL: http://asset.jmir.pub/assets/f8f61fedd62854c16e093af491ceb9e0.pdf