Journal of Multidisciplinary Healthcare

Dovepress

Open Access Full Text Article

ORIGINAL RESEARCH

Precision Structuring of Free-Text Surgical Record for Enhanced Stroke Management: A Comparative Evaluation of Large Language Models

Mengfei Wang[1], Jianyong Wei[2], Yao Zeng[3], Lisong Dai[4], Bicong Yan[4], Yueqi Zhu[4], Xiaoer Wei[4], Yidong Jin[1], Yuehua Li[4]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, People's Republic of China; [2]Clinical Research Center, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China; [3]Shenyang University of Technology, Shenyang, Liaoning, People's Republic of China; [4]Institute of Diagnostic and Interventional Radiology, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China

Correspondence: Yuehua Li, Institute of Diagnostic and Interventional Radiology, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, No. 600 Yishan Road, Shanghai, 200233, People's Republic of China, Email liyuehua0529@163.com

Introduction: Mechanical thrombectomy (MTB) is a critical procedure for acute ischemic stroke (AIS) patients. However, the free-text format of MTB surgical records limits the formulation of effective postoperative patient management and rehabilitation plans. This study compares the efficacy of large language models (LLMs) in structuring data from these free-text MTB surgical record.

Methods: This retrospective study collected a total of 382 MTB surgical records from a tertiary hospital. An initial analysis of 30 surgical record from these records provided a guiding prompt for LLMs, focusing on basic and advanced characteristics, such as occlusion locations, thrombectomy maneuvers, reperfusion status, and intraoperative complications. Six LLMs—ChatGPT, GPT-4, GeminiPro, ChatGLM4, Spark3, and QwenMax—were assessed against data extracted by neuroradiologists and a junior physician for comparison. The all 382 surgical records were used to test the performance of LLMs. The performance of the LLMs was quantified using Accuracy, Sensitivity, Specificity, AUC, and MSE as an additional metric for advanced characteristics.

Results: All LLMs showed high performance in characteristic extraction, achieving an average accuracy of 95.09 ± 4.98% across 48 items, and 78.05 ± 4.2% overall. GLM4 and GPT-4 were most accurate in advanced characteristics extraction, with accuracies of 84.03% and 82.20%, respectively. The processing time for LLMs averaged 73.10 ± 10.86 seconds of six models, significantly faster than the 427.88 seconds for manual extraction by physicians.

Conclusion: LLMs, particularly GLM4 and GPT-4, efficiently and accurately structured both general and advanced characteristics from MTB surgical record, outperforming manual extraction methods and demonstrating potential for enhancing clinical data management in AIS treatment.

Keywords: large language models, free-text report, mechanical thrombectomy, acute ischemia stroke

Introduction

Acute Ischemic Stroke (AIS) stands as a leading cause of mortality and debilitating conditions among adults globally.[1] The burden of stroke has increased significantly over recent decades, with the Global Burden of Disease (GBD) study showing that the annual number of strokes and stroke-related deaths increased substantially from 1990 to 2019.[2] This trend is particularly evident in China, where the burden of stroke remains critically high.[3,4]

The surgical records of mechanical thrombectomy (MTB) detail essential intraoperative findings, including the extent of blood flow restoration and reperfusion status. These records are crucial for developing personalized, systematic, and evidence-based post-surgical management strategies, ultimately improving patient outcomes and quality of life.[5]

The prevalent use of free-text for documenting mechanical thrombectomy introduces inefficiencies in medical data utilization. Extracting precise data from such texts is error-prone and incomplete, complicating the integration with

Received: 10 July 2024
Accepted: 18 October 2024
Published: 14 November 2024

structured systems like Electronic Health Records.[6,7] This disjunction not only stymies extensive clinical research but also hinders clinical decision-making and effective patient follow-up, as the information is poorly suited for quick reference or integration into clinical decision support systems. Adopting structured reports for mechanical thrombectomy operations in AIS patients holds considerable clinical benefits. Structured reports enhance the speed and clarity of presenting surgical data, aiding healthcare providers in making informed and timely treatment decisions. Integration of structured reports into clinical decision support systems bolsters the accuracy and scientific basis of treatment strategies. Furthermore, organized documentation is crucial for monitoring patient recovery and refining follow-up protocols, thereby elevating both clinical efficiency and patient care standards.[8]

Large language models (LLMs) such as ChatGPT,[9] GPT4,[10] and Gemini have performed well on various downstream natural language processing(NLP) tasks and are widely used in the medical field.[11] Previous studies have evaluated LLMs' capabilities in extracting structured imaging reports[12,13] and in medical professional exams.[14,15] However, these studies have primarily focused on English texts, using data mainly from open-source datasets or specific public examination questions, and predominantly on ChatGPT or GPT-4. Given the complexity and unpredictability of surgery, the content and organization of free-text surgical records exhibit much greater variability and heterogeneity compared to semi-standardized electronic medical records or radiology reports. This variability can pose challenges in extracting information from these documents.

To further enhance the potential of MTB surgical records in formulating effective postoperative patient management and rehabilitation plans for AIS patients, this study aims to perform a comparative evaluation of the capabilities of six widely used large language models (LLMs). These include three globally recognized models: ChatGPT (GPT-3.5 Turbo), GPT-4, and Gemini Pro,[16] as well as three major Chinese models: ChatGLM-4 (GLM4),[17] Spark 3,[18] and Qwen Max, in extracting vital information from free-text Chinese MTB surgical records and generating structured text. It is well-known that GPT-4 performs exceptionally well in many tasks. In our setup for extracting text from Chinese MTB records, we are interested in whether GPT-4 can maintain its strong performance. Additionally, we are keen to see if other large language models, especially those developed in China, can outperform GPT-4 in our specific context.

Materials and Methods

Dataset

A total of 502 consecutive reports from patients with AIS who underwent MTB treatment at Shanghai Sixth People's Hospital from September 2019 to November 2023 were retrospectively collected from the hospital information system between January 3, 2024, and January 12, 2024. These reports meticulously documented various examinations and treatment processes during hospitalization, including detailed records of the MTB treatment procedure, all in free-text format. Inclusion criteria included: (1) patient age over 18 years; (2) AIS resulting from CT-confirmed large vessel occlusion of the middle cerebral artery; (3) treatment with MTB. The exclusion criterion was the absence of a detailed report (n = 120 patients). A total of 382 patients were finally included. To better illustrate the enrollment process, a flow chart has been provided (Figure 1). Before analysis, all personal data were anonymized, resulting in 382 MTB surgical record with free-text descriptions of the MTB procedure for further research in Chinese. There was neither an internal nor an external data set.

Using of the Large Language Models

This study included six large language models capable of supporting simplified Chinese queries, including: (1) GPT-3.5 turbo, which has a parameter count in the hundreds of billions and has garnered widespread global attention.[19] (2) GPT-4, the successor to GPT-3.5 turbo, boasts an even larger parameter count and has demonstrated outstanding performance in multiple medical scenario tests.[14,20] (3) GLM4, which has matched or approached GPT-4's performance in several benchmark tests, particularly in Chinese understanding.[17] (4) Spark3, which has undergone specialized training in multiple professional fields, including medicine, resulting in significantly improved performance.[18] (5) QwenMax and GeminiPro were also included in this study to enrich the baseline for experimental comparison. The information about the development companies and countries of the above models can be found in the Table S1.

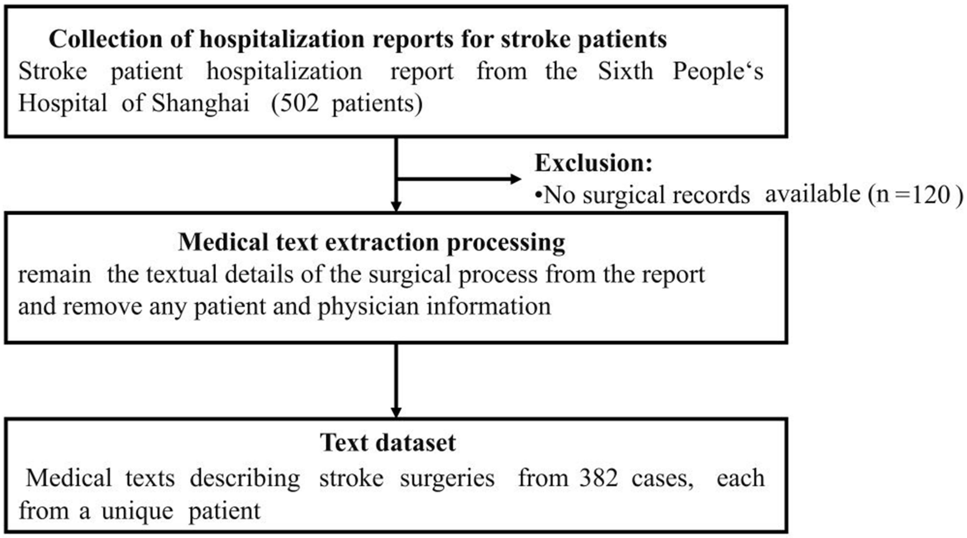
Collection of hospitalization reports for stroke patients
Stroke patient hospitalization report from the Sixth People‘s Hospital of Shanghai (502 patients)
Exclusion:
•No surgical records available (n =120)
Medical text extraction processing
remain the textual details of the surgical process from the report and remove any patient and physician information
Text dataset
Medical texts describing stroke surgeries from 382 cases, each from a unique patient

Figure 1 Patient Enrollment Flow Chart. Flow chart depicting the process of patient enrollment and exclusion for the study, including the initial cohort of 502 patients, the exclusion criteria, and the final number of 382 patients included in the analysis.

To ensure accuracy and reproducibility, all models were accessed via their official default interfaces. Python programs were used to implement cyclic question calling, with no modifications to the models' default hyperparameters during the process. To explore the application capabilities of LLMs in medical text processing, this study designed two experiments: one focused on extracting key basic characteristics, and the other on generating advanced logic characteristics from MTB procedure details. An initial analysis of 30 surgical record from these records provided a guiding prompt for LLMs. The prompts for each experiment were constructed by a senior neuroradiologist with over 15 years of experience (detailed in Supplementary Material 1). Characteristics were extracted by LLMs from free-text MTB surgical records and returned in a structured JSON format. In cases where the JSON format was incorrect, we re-requested the specific case to obtain a properly formatted response. A post-processing tool was developed to transfer all model-generated JSON texts to corresponding Excel spreadsheets for further analysis. This workflow is illustrated in Figure 2.

Key Basic Characteristics Extraction from MTB Surgical Record

The operating records in MTB procedure report meticulously noted the thrombolytic drugs used, the exact occlusion locations, the TICI grading, and whether there were any occurrences of bleeding or hematoma post-surgery. This experiment aimed to test the ability of six LLMs to extract data from MTB surgical record and return key characteristics in a structured JSON format. A prompt was tested on 30 records to identify errors and optimize instructions for the LLMs. The experiments were conducted in Chinese, with an English version of the prompt provided in Supplementary Material 1. This experiment was based on 382 detailed free-text records describing MTB procedures, comprehensively covering essential information such as occlusion sites, procedural processing, angiographic outcomes, therapeutic measures, and treatment outcomes. Specific examples are provided in Supplementary Material 1 and Figure S1. The characteristics that were extracted, as well as the instructions for the LLMs for each procedural detail, are summarized in Table 1. These characteristics are crucial for assessing patient functional outcomes and serve as important indicators that assist in clinical decision-making.[21]

For reference standard construction, all records were independently evaluated by two interventional neuroradiologists (X.X.X and X.X.X, with over 5 years of experience), and disagreements were resolved by an arbitrator (X.X.X., with 15 years of experience in neurointerventional surgery). During this process, the neuroradiologists were blinded to the results produced by the LLMs. The data were exclusively extracted from MTB surgical records, with no data acquired from other clinical sources.

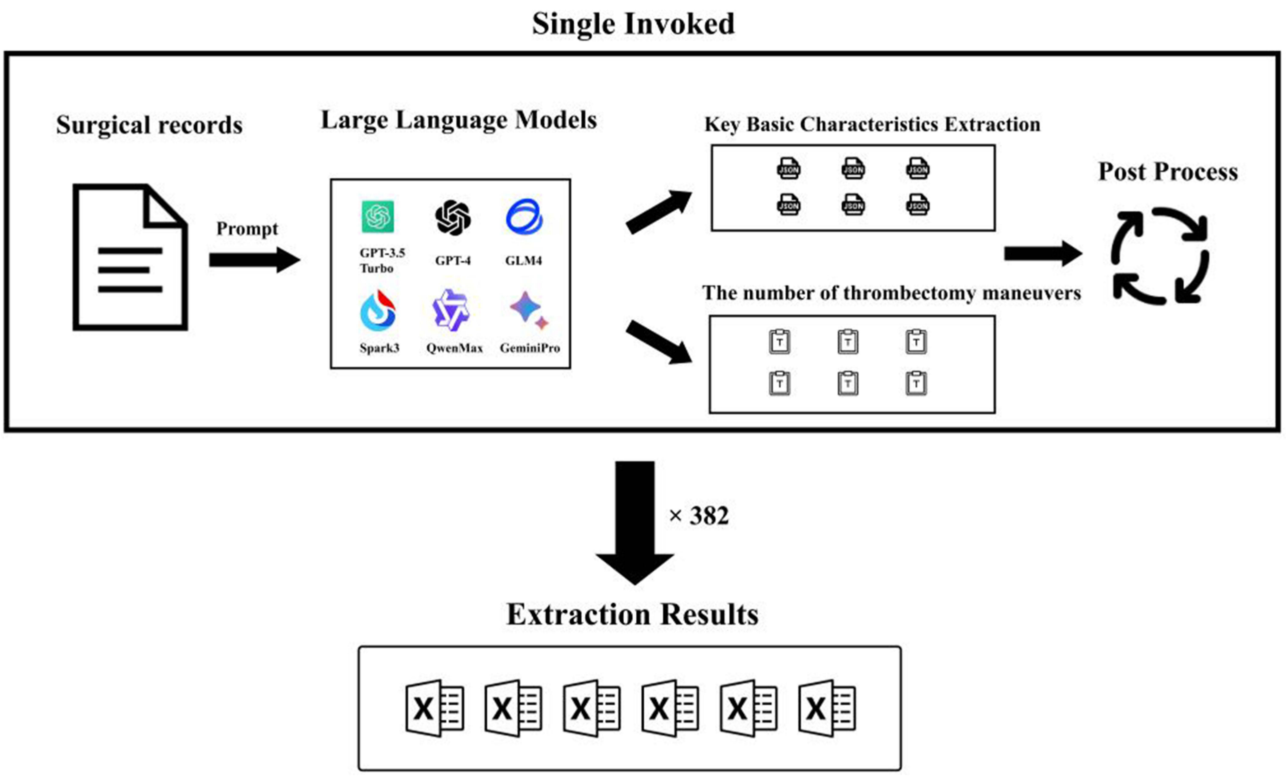

Figure 2 Large Language Model Invocation and Data Retrieval Process. Each Chinese medical text is input into six large language models, which analyze the text and return JSON outputs (for retrieving structured item information) or text regarding the number of thrombectomy maneuver. After relevant text post-processing, all data are transferred and saved into an xlsx file.

Advance Logic Characteristics from MBT Surgical Record

This experiment focused on extracting a characteristic generated by advanced logic using LLMs. The models were tasked with calculating the number of thrombectomy maneuvers in MTB surgical record, a task that requires both medical knowledge and logical reasoning capabilities. Typically, the number of thrombectomy maneuvers may not be directly recorded as a number in the records but described in text, such as indicating statements like “The microcatheter was then guided again by the microguidewire to the distal end of the right M1 segment” or “Repeat the above procedure three times”. This experiment involved the textual analysis of the same set of 382 detailed medical texts documenting MTB

Table 1 Summary of Instructions and Options Given to the Large Language Models

Extracted Key characteristics	Instructions
Thrombectomy performing	Yes/No
Stent using	Yes/No
Location of vessel occlusion	The naming of segment vessel, such as left MCA, right MCA, right ICA, left ICA, BA, or unknown
Balloon guide catheter using	Yes/No
Midline shift of brain	Yes/No
Intravenous thrombolysis	Yes/No
Bleeding or hematoma	Yes/No
mTICI grading	Grading value (eg, 0, 1, 2a, 2b, 3) / Unknown

Abbreviations: mTICI, modified Thrombolysis in Cerebral Infarction; MCA, middle cerebral artery; ICA, internal carotid artery; BA, basilar artery.

treatment. The first pass effect was defined as the scenario explicitly indicated in the text where a guidewire was introduced into the vessel, a single pass of the device achieved complete revascularization of the large vessel occlusion and its downstream territory (mTICI 3), and no rescue therapy was employed. The six LLMs (invoked on January 22, 2024) analyzed all 382 records using the same Chinese prompt to extract the number of thrombectomy maneuvers. For comparison, a junior neuroradiologist (XXX, with 3 years of experience) independently analyzed all 382 records to extract the number of thrombectomy maneuvers. An expert with over 15 years of experience in neurointerventional surgery analyzed all the records and these findings served as the reference standard.

Statistical Analysis

All results were numerically transformed before statistical analysis. Only data entries from the LLMs that exactly matched the expert's readings were counted as correct. Any deviation from the given options in the prompt, including synonyms, punctuation marks, or any additional symbols entered by the LLM, was counted as false. If a data point was not included in the report, it was declared as "missing" or "unknown" by the neuroradiologist. When the LLM also declared certain information as "missing" or "unknown", the data entry was categorized as correct. The performance of these models and junior neuroradiologists was assessed by calculating accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC). The comparison of AUC values between groups was conducted using the DeLong test.[22,23] The significance of differences in accuracy between the model and junior doctors was assessed using the McNemar test. For the occlusion locations and TICI grading, only accuracy was calculated. The 95% confidence intervals were determined by bootstrapping with 1000 samples. Statistical analyses were performed by one of the authors (X.X.X.) using Python 3.10. The total number of tokens in the text was calculated using https://github.com/openai/tiktoken.

Results

Patient Characteristics

A total of 382 MTB surgical records from 382 patients (mean age, 72.23 years ± 13.35) were finally included for analysis, comprising 208 male (208/382, 54.45%) and 174 female (174/382, 45.55%). The average character length of the surgical records in report was 277.21 characters, with an average token count of 588.88. Texts with a mTICI score of 3 accounted for 53.79%, while texts documenting occlusions of the middle cerebral artery comprised 63.61%. There were 333 texts involving thrombectomy procedures, 310 texts detailing stent usage, 80 texts on balloon dilation, 83 texts on intravenous thrombolysis, one text describing midline shift, and 17 texts reporting bleeding or hematoma. Patient characteristics and basic clinical and procedural details are summarized in Table 2

Key Basic Characteristics Extraction by LLMs

Performance of Models in Key Information Extraction

In the process of extracting basic key information from 382 texts of 382 patients, the models performed excellently in most key information extraction tasks, with an average accuracy of 95.09 ± 4.98% in 8 items*8 models = 64 items. Specifically, GPT-3.5 Turbo achieved an average accuracy of 92 ± 6.72%; Gemini Pro had an average accuracy of 94.18 ± 4.57%; GLM4's average accuracy was 95.91 ± 3.75%; GPT-4 reached an average accuracy of 97.67 ± 2.60%; Qwen Max had an average accuracy of 95.02 ± 5.12%; and Spark 3's accuracy was 95.58±4.34%. Figure 3 shows the performance accuracy of six large language models with a line chart.

For the six key information extraction tasks, With the "mTICI score" and "occlusion location" excluded, GPT-3.5 Turbo's average AUC was 0.91 ± 0.09, Gemini Pro's average AUC was 0.90 ± 0.008, GLM4's average AUC was 0.97 ± 0.04, GPT-4's average AUC was 0.92 ± 0.08, Qwen Max's average AUC was 0.96 ± 0.05, and Spark 3's average AUC was 0.95 ± 0.04. More details can be found in Table 3, and the Delong Test and additional comparison details are provided in Tables S2–S4.

Table 2 Patient Characteristics and Key Procedural Information from Text

Characteristic	Statistic
Sex	
F	174(45.55)
M	208(54.45)
Mean age (y)*	72.2 ± 13.4
Age range (y)[†]	22–96(74)
Thrombectomy procedures	333(87.17)
Stent used	310(81.15)
Balloon guide catheter	80(20.94)
Intravenous thrombolysis	83(21.17)
Midline shift	1(0.26)
Bleeding or hematoma	16(4.18)
mTICI	
0	1 (0.26)
1	4 (1.05)
2a	7 (1.83)
2b	23 (6.02)
2c	7 (1.83)
3	206 (53.93)
Unknown	129(33.77)
Occlusion site	
MCA	243 (63.61)
BA	48 (12.57)
ICA	3 (0.79)
ACA	32 (8.38)

Notes: Unless otherwise indicated, data are numbers of patients and data in parentheses are percentages. *Data are mean ± SDs. [†]Data in parentheses are the range(median).

Incorrect Data Analysis

In an experiment involving the extraction of basic key information from texts of 382 patients, GPT-4 and Spark 3 were the only models that consistently returned data in the correct format, while other models exhibited varying degrees of format errors. Specifically, GPT-3.5 Turbo and GLM4 exhibited format error rates of 0.52% (2 cases) and 0.79% (3 cases), respectively. In contrast, both Gemini Pro and Qwen Max had higher error rates of 8.2% (34 cases each).

For the extraction of occlusion locations, GPT-4 demonstrated an accuracy of 98.41% [95% CI: 98.03, 99.02], and GLM4 recorded an accuracy of 96.58% [95% CI: 95.74, 97.38], both outperforming other models. Some patients had multiple occlusion sites, and the original texts were often complex. Some patients presented with multiple occlusion sites,

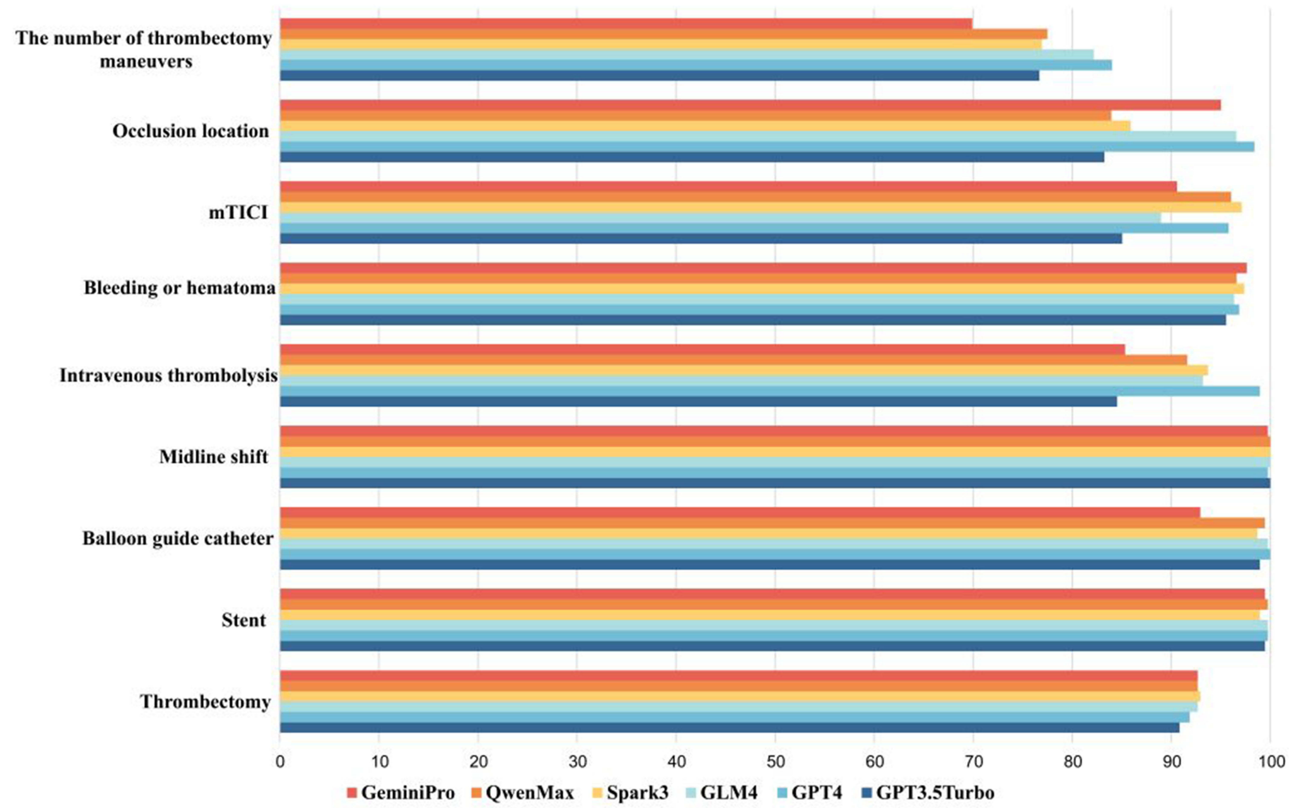

Figure 3 A line chart showing the accuracy of each model across various items. The performance accuracy of six large language models—GeminiPro, QwenMax, Spark3, GLM4, GPT-4, and GPT-3.5 Turbo—across different clinical information extraction categories from mechanical thrombectomy surgical records. The evaluated tasks include the number of thrombectomy maneuvers, occlusion location identification, mTICI grading, detection of bleeding or hematoma, intravenous thrombolysis, midline shift, usage of a balloon guide catheter, stent usage, and identification of thrombectomy procedures. Each bar represents the accuracy percentage for each model, highlighting comparative effectiveness in the extraction tasks.

complicating the extraction process. Certain models frequently overlooked parts of the occluded vessels or specific occlusion locations. For instance, instead of the expected "Left MCA", a model might inaccurately respond with "MCA M1".

In the extraction of mTICI scores, Spark 3.0 had an accuracy of 97.11% [95% CI: 96.39, 98.03]. Other models, such as GLM4 and GPT-3.5 Turbo, might deduce an mTICI score from the text information even when the original text lacked

Table 3 Performance of Each Model in Basic Key Information Extraction

	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Thrombectomy	GPT3.5Turbo	90.84 (87.95–93.72)	91.16 (88.17–94.13)	85.00 (66.67–100.00)	88.08 (78.82–95.62)
	GeminiPro	92.67 (89.79–95.03)	93.20 (90.39–95.67)	**86.21** (72.00–96.67)	**89.70** (82.37–95.41)
	GLM4	92.67 (90.05–95.29)	93.95 (91.40–96.32)	80.00 (66.67–91.67)	86.97 (79.98–92.74)
	GPT4	91.88 (88.74–94.50)	92.42 (89.52–95.17)	84.62 (68.75–96.43)	88.52 (80.41–94.79)
	QwenMax	92.67 (90.05–95.03)	93.95 (91.44–96.32)	80.00 (65.79–91.44)	86.97 (79.80–92.97)
	Spark3	**92.93** (90.31–95.29)	**94.74** (92.33–96.95)	77.50 (64.00–89.13)	86.12 (79.27–92.00)
Stent	GPT3.5Turbo	99.48 (98.69–100.00)	99.36 (98.37–100.00)	**100.00** (100.00–100.00)	99.68 (99.19–100.00)
	GeminiPro	99.48 (98.69–100.00)	**99.68** (98.98–100.00)	98.61 (95.58–100.00)	99.14 (97.50–100.00)
	GLM4	**99.74** (99.21–100.00)	**99.68** (98.99–100.00)	**100.00** (100.00–100.00)	**99.84** (99.49–100.00)
	GPT4	**99.74** (99.21–100.00)	**99.68** (98.99–100.00)	**100.00** (100.00–100.00)	**99.84** (99.49–100.00)
	QwenMax	99.74 (99.21–100.00)	**99.68** (98.99–100.00)	**100.00** (100.00–100.00)	**99.84** (99.49–100.00)
	Spark3	98.95 (97.91–99.74)	98.73 (97.39–99.69)	**100.00** (100.00–100.00)	99.36 (98.70–99.85)

(*Continued*)

Table 3 (Continued).

	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Balloon guide catheter	GPT3.5Turbo	98.95 (97.90–99.74)	96.34 (92.10–100.00)	99.67 (98.95–100.00)	98.00 (95.78–99.68)
	GeminiPro	92.93 (90.31–95.29)	75.24 (66.98–83.49)	99.64 (98.87–100.00)	87.44 (83.16–91.40)
	GLM4	99.74 (99.21–100.00)	**100.00** (100.00–100.00)	99.67 (98.97–100.00)	99.83 (99.48–100.00)
	GPT4	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)
	QwenMax	99.48 (98.69–100.00)	98.75 (95.65–100.00)	99.67 (98.97–100.00)	99.21 (97.62–100.00)
	Spark3	98.69 (97.38–99.74)	95.18 (90.11–98.91)	99.67 (98.96–100.00)	97.42 (94.85–99.45)
Midline shift	GPT3.5Turbo	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)
	GeminiPro	99.74 (99.21–100.00)	50.00 (0.00–100.00)	**100.00** (100.00–100.00)	75.00 (50.00–100.00)
	GLM4	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)
	GPT4	99.74 (98.95–100.00)	50.00 (0.00–100.00)	**100.00** (100.00–100.00)	75.00 (50.00–100.00)
	QwenMax	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)
	Spark3	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)	**100.00** (100.00–100.00)
Intravenous thrombolysis	GPT3.5Turbo	84.55 (81.15–87.96)	96.15 (87.49–100.00)	83.71 (80.00–87.22)	89.93 (85.12–93.26)
	GeminiPro	85.34 (82.20–88.48)	**100.00** (100.00–100.00)	84.23 (80.85–87.43)	92.11 (90.42–93.71)
	GLM4	93.19 (90.58–95.55)	**100.00** (100.00–100.00)	92.00 (89.19–94.80)	96.00 (94.59–97.40)
	GPT4	**98.95** (97.91–99.74)	**100.00** (100.00–100.00)	**98.68** (97.28–99.68)	**99.34** (98.64–99.84)
	QwenMax	91.62 (88.74–93.98)	**100.00** (100.00–100.00)	90.33 (87.12–93.06)	95.17 (93.56–96.53)
	Spark3	93.72 (91.35–95.81)	**100.00** (100.00–100.00)	92.57 (89.67–95.00)	96.28 (94.83–97.50)
Bleeding or hematoma	GPT3.5Turbo	95.55 (93.46–97.38)	50.00 (12.50–87.50)	96.52 (94.41–98.38)	73.26 (54.65–91.76)
	GeminiPro	**97.64** (96.07–98.96)	**100.00** (100.00–100.00)	**97.59** (95.93–98.95)	**98.80** (97.97–99.47)
	GLM4	96.34 (94.50–98.17)	**100.00** (100.00–100.00)	96.31 (94.43–98.16)	98.15 (97.21–99.08)
	GPT4	96.86 (95.03–98.43)	85.71 (50.00–100.00)	97.07 (95.17–98.67)	91.39 (73.68–99.20)
	QwenMax	96.60 (94.76–98.43)	**100.00** (100.00–100.00)	96.56 (94.69–98.41)	98.28 (97.35–99.20)
	Spark3	97.38 (95.55–98.95)	88.89 (60.00–100.00)	**97.59** (95.74–98.94)	93.24 (78.67–99.33)
mTICI	GPT3.5Turbo	85.06 (83.28,86.89)			
	GeminiPro	90.57 (89.18,92.13)			
	GLM4	88.99 (87.54,90.49)			
	GPT4	95.79 (94.75,96.72)			
	QwenMax	96.06 (95.08,97.05)			
	Spark3	**97.11** (96.39,98.03)			
Occlusion location	GPT3.5Turbo	83.27 (81.63,85.25)			
	GeminiPro	95.04 (94.10,96.39)			
	GLM4	96.58 (95.74,97.38)			
	GPT4	**98.41** (98.03,99.02)			
	QwenMax	83.96 (81.02,84.95)			
	Spark3	85.88 (84.26,87.55)			

Notes: Unless otherwise indicated, data are percentages and data in parentheses are 95% CIs. The bold text represents the highest value of a single indicator within a project.

a specific mTICI score. Another area where models generally showed poor performance was in determining whether MTB had been performed. This task likely requires extensive medical knowledge and sophisticated logical reasoning, presenting a significant challenge for LLMs. This also underpins the rationale behind designing the Medical Logic Analysis experiment.

Medical Logic Analysis

Thrombectomy Count Extraction

In this study, we analyzed the number of thrombectomy procedures mentioned in 382 medical texts related to stroke surgery. This task evaluated not only the clinical medical knowledge of each model but also their capabilities in comprehensive medical analysis and reasoning. Table 4 shows the accuracy and mean squared error (MSE) of six models on thrombectomy count extraction. The average accuracy of the six models across eight items was 78.05±4.2%. Among them, the GLM4 and GPT-4 models demonstrated slightly higher accuracies of 84.03% [95% CI: 80.37, 87.70]

Table 4 Accuracy and Mean Squared Error Statistics for Thrombectomy Count by Each Model

Models	Accuracy (%)	MSE
GPT3.5Turbo	76.70(72.25,80.89)	0.32(0.24,0.41)
GeminiPro	69.90(65.45,74.35)*	2.04(0.96,3.63)
GPT4	82.20(78.27,85.87)	0.31(0.18,0.48)
GLM4	84.03(80.37,87.70)*	0.24(0.15,0.35)
Spark3	77.49(73.56,81.68)	0.70(0.30,1.20)
QwenMax	76.96(72.77,80.90)	0.47(0.32,0.63)
Junior Physician	79.06(74.87,82.99)	0.33(0.24,0.43)

Notes: Unless otherwise indicated, data are percentages and data in parentheses are 95% CIs. Comparisons between models and Junior Physician were made using the McNemar test. *P<0.05.

and 82.20% [95% CI: 78.27, 85.87] respectively, surpassing the other models (p<0.05). In comparison, a junior physician achieved an accuracy of 79.06% [95% CI: 74.87, 82.99]. Notably, the GeminiPro model performed less effectively than the junior physician (p < 0.05).

Differentiation of Initial Thrombectomy Recanalization

In distinguishing the initial thrombectomy recanalization across 382 MTB surgical records, the average AUC for all models was 0.80±0.03 (Table 5). Both GPT4 (AUC=0.85 [95% CI: 0.82 −0.88]) and GLM4 (AUC=0.85[95% CI: 0.82 −0.89]) outperformed other models (p<0.05) and were statistically similar to the junior physician (AUC=0.86 [95% 0.82 −0.90]), demonstrating superior performance. The comparison of GLM4 and GPT4 by Junior Physician is shown in the Bland-Altman plot in Figure 4.

To conclude our results, we compared the performance of various models in predicting clinical outcomes related to thrombectomy procedures. Figure 5 provides a visual representation of the accuracy of six models—GPT-3.5 Turbo, GPT-4, GLM4, Spark3, QwenMax, and GeminiPro—across key metrics relevant to interventional success. These metrics include thrombectomy success rate, stent usage, balloon guide catheter deployment, occlusion location identification, mTICI scores, the incidence of bleeding or hematoma, and the number of thrombectomy maneuvers required.

Discussion

This study performed a comprehensive evaluation of six LLMs using real clinical EMR data to extract essential characteristics from free-text records of MTB treatment and to generate structured text in Chinese. A comparative study was conducted to explore the differences in basic key characteristics extraction capabilities among the different models (Figures 3 and 5). Additionally, the performance of six LLMs was compared with that of physicians in advanced logic generated characteristics from text reports. All six LLMs performed well in free-text information extraction, with

Table 5 Performance Statistics of Each Model in Distinguishing First Thrombectomy Maneuver

	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
GPT3.5Turbo	79.40 (75.39 −83.25)	83.36 (78.00 −87.92)	74.68 (67.74 −81.32)	78.90 (74.62 −83.13)*
GeminiPro	74.75 (70.42 −79.06)	72.90 (66.66 −78.68)	77.46 (71.13 −83.43)	74.97 (70.61 −79.31)*
GPT4	84.52 (80.63 −87.96)	76.98 (70.91 −82.59)	93.16 (89.56 −96.63)	85.12 (81.55 −88.38)
GLM4	85.95 (82.20 −89.27)	95.12 (91.96 −97.71)	75.08 (68.89 −81.25)	85.18 (81.50 −88.66)
Spark3	81.17 (77.23 −85.08)	97.55 (95.43 −99.49)	61.98 (54.70 −69.02)	79.94 (75.94 −83.90)*
QwenMax	80.25 (76.44 −84.04)	72.30 (66.33 −78.42)	89.85 (85.14 −94.27)	81.05 (77.41 −84.76)*
JuniorPhysician	86.25 (82.72 −89.53)	87.82 (83.42 −92.17)	84.06 (78.61 −89.54)	86.01 (82.30 −89.51)

Notes: Unless otherwise indicated, data are percentages and data in parentheses are 95% CIs. Comparisons between models and Junior Physician were made using the Delong test. *P<0.05, indicating a significant difference compared to Junior Physicians.

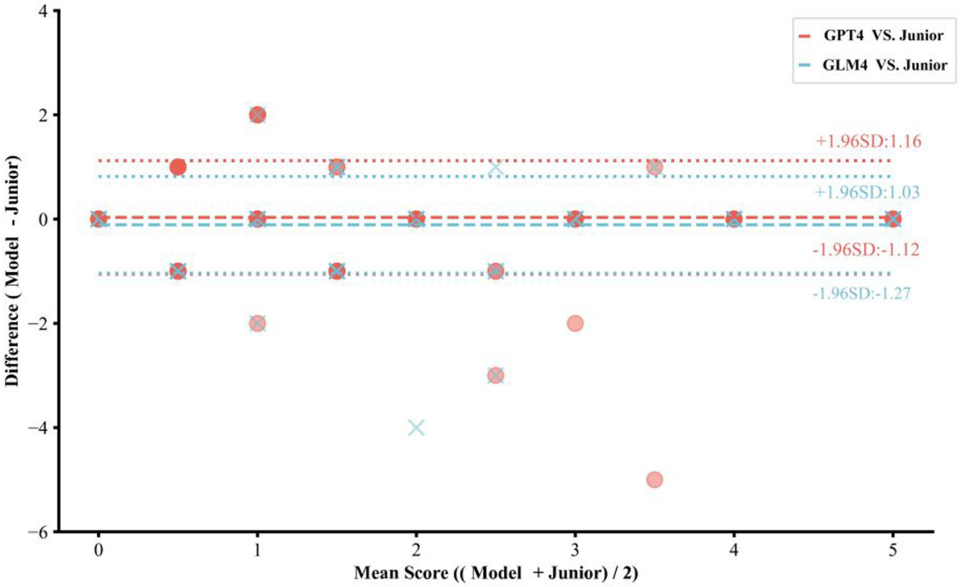

Figure 4 Comparison of the differences in the number of thrombectomy maneuver extracted between GPT-4, GLM-4, and Junior Physician. The agreement between the number of thrombectomy maneuvers extracted by GPT-4 and GLM4 models compared to a junior physician. The mean difference between each model and the junior physician is plotted against the average score, with dotted lines indicating the limits of agreement (±1.96 standard deviations). The red markers and dashed lines represent GPT-4 vs Junior Physician, while the blue markers and dashed lines represent GLM4 vs Junior Physician.

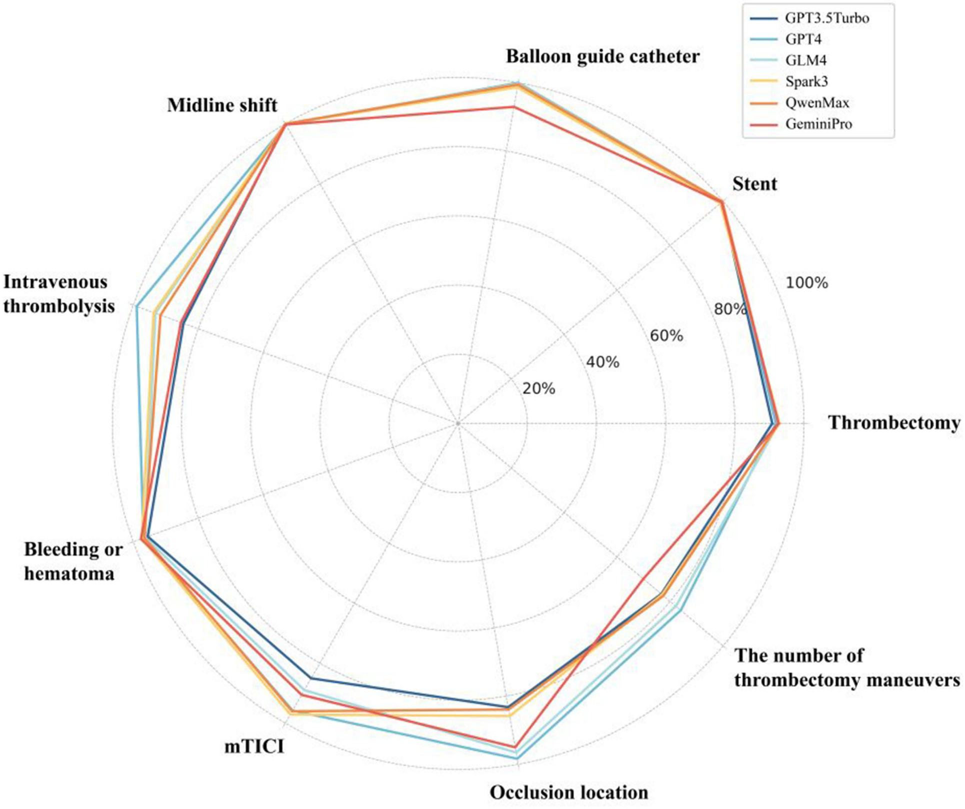

Figure 5 This is a radar chart showing the accuracy of each model across various items. The metrics represented include thrombectomy, stent using, balloon dilation, occlusion location, mTICI scores, bleeding or hematoma and intravenous thrombolysis the number of thrombectomy maneuver. Each line represents a different model. The area covered by each line corresponds to the model's performance across these categories, as detailed in the following: GPT-3.5 Turbo = 23,742.56, GeminiPro = 24,152.32, GLM4 = 25,771.35, GPT-4 = 26,759.46, QwenMax = 25,119.24, Spark3 = 25,353.42.

an accuracy of 92% to 97.67% for extracting key information. For advanced characteristics, the AUCs for extracting the number of thrombectomy maneuvers ranged between 0.90 and 0.97 (Table 3 and Figure 3). In the advanced logic generated characteristics extraction of accurately extracting the number of thrombectomy maneuvers, GLM4 and GPT-4 outperformed the other models (all p<0.05), and GLM4 performed better than junior Physician (p<0.05). Regarding the identification of the first-pass effect, GLM4 and GPT-4 performed similarly to a junior physician (p=0.66 and p=0.67) and better than other models (p<0.05).

Large language models based on deep learning and Transformers show great potential in extracting critical clinical information.[11] Recently, numerous studies have utilized GPT-4 and ChatGPT to extract key information from lung cancer CT imaging reports.[24] GPT-4 achieved an accuracy of 96%, while ChatGPT achieved an accuracy of 93.7% in another task involving Chinese text-based imaging reports.[25] These studies demonstrate the excellent capability of ChatGPT and GPT-4 in extracting information from imaging reports. The EMR texts contain vast amounts of information pertinent to clinical research; extracting these key details is crucial for advancing clinical studies and translation, as well as improving clinical outcomes (5). For instance, the number of thrombectomy maneuvers and the success of the first-pass in MTB treatments are essential for assessing treatment efficacy and function outcome. Studies indicate that successful thrombectomy in first-pass effect can significantly enhance clinical efficacy and yield better clinical outcomes,[26] with successful reperfusion on the first attempt being a critical condition for the first-pass effect in MTB. Recently, Nils et al tested the accuracy of GPT-4 in extracting key information from 100 German MTB reports,[27] achieving an accuracy of 94%, compared to 63% for GPT-3.5. Our study compared six LLMs on 382 real Chinese MTB surgical record from a hospital setting, where GPT-4 achieved an average accuracy of 95.95% across nine tasks (key information extraction + thrombectomy count extraction), and GPT-3.5 Turbo achieved an average accuracy of 90.04%. GPT-4's performance was similar to the results reported by Nils et al. In the key information extraction tasks, we also calculated the sensitivity, specificity, and AUC values of each model and used the DeLong test to comprehensively assess the differences in various metrics among the models. Additionally, due to differences in language and report types, the number of evaluated categories in our study was fewer than that in Nils et al's research. However, we specifically evaluated the models' ability to extract the number of thrombectomy maneuvers, a task requiring significant medical logic, and compared their performance with a junior physician. For this task, GPT-4 and GLM4 achieved AUC values comparable to the junior physician (p>0.05). In our study, the three models from China including GLM4, QwenMax and Spark3 demonstrated commendable performance. This may be due to their specialized optimization for Chinese and medical contexts.[16,17] In structured text extraction tasks, even without medical terminology explained in the prompts, GPT-4 and GLM4 managed to extract specified items from the reports with a high performance with accuracy of 95.91% and 97.67% in the basic key characteristics extraction, such as "mTICI score", "hemorrhage and hematoma", and "occlusion location".

Key information from medical texts like stroke surgery records is often used to train machine learning models for downstream tasks, thus developing and proposing more advanced clinical diagnostic and disease assessment models.[20] Without specific tools for particular text processing tasks, using large language models to extract some key information can speed up workflows, reduce workload, and promote analyses related to stroke prognosis and treatment. Studies have shown that GPT-4 performs comparably to existing professional models in structured tasks involving imaging reports.[23] As the cost of large models decreases, the expense of processing medical texts with them will also be further reduced. For example, using GPT-4, at the cost of $0.006 / 1K tokens, we processed 382 texts, totaling 226k tokens, which costs approximately $1.36 according to OpenAI's rate as of January 22, 2024.

Our study has several limitations. First, our study primarily explored the performance of various models on Chinese medical texts. The prompts used in this study might not be suitable for all models, although they still exhibited excellent performance with high accuracy. Secondly, retrospective data collection would lead to class imbalance, such as midline shift and hemorrhage, which are rare among AIS patients. In our study of 382 patients, only one presented midline shift, and 17 cases had hemorrhages. Therefore, we employed the AUC as a metric to minimize the impact of class imbalance on the final results. Third, the sample size in this study was limited, necessitating additional data and experiments to more comprehensively evaluate the performance of LLMs in medical text processing.

Conclusion

This study demonstrated the effectiveness and potential of LLMs in processing free-text Chinese medical reports, particularly in extracting key information from mechanical thrombectomy treatment records. The results indicate significant differences among various LLMs in handling complex medical texts, with GLM4 and GPT-4 showing superior performance in most evaluated tasks. Specifically, these models either matched or surpassed the accuracy of junior physicians in advanced logic extraction tasks, achieving an average accuracy of over 82%. However, despite the promising results, re-evaluation by more experienced physicians is still necessary to ensure the clinical reliability and safety of the extracted information.

Abbreviations

AIS, Acute ischemia stroke; MTB, Mechanical thrombectomy; LLM, Large language model.

Data Sharing Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

This study has obtained approval from the Ethics Committee of Shanghai Sixth People's Hospital [IRB code: 2024-061] and has been conducted in compliance with the Declaration of Helsinki. The Ethics Committee has waived the requirement for individual informed consent due to the retrospective nature of the study. All patient data were anonymized, and confidentiality was strictly maintained throughout the research process.

Acknowledgments

The authors used large language models GPT-4, GPT-3.5 (OpenAI; https://chat.openai.com), GeminiPro (Google: https://deepmind.google/technologies/gemini/pro/), ChatGLM4 (Zhipu: https://chatglm.cn), QwenMax (AlibabaCloud: https://tongyi.aliyun.com/qianwen/) and Spark3(Iflytech: https://xinghuo.xfyun.cn/) to process the Surgical Record. Zhuo Li from Shenyang University of Technology assisted in the typesetting of the article.

Funding

This research was supported by the National Natural Science Foundation of China Grant No. 8225024 and No. 81871329. Funding agencies did not play a role in study design, data collection, analysis and interpretation, and manuscript writing." Revised text: "This research was supported by the National Natural Science Foundation of China Grant No. 8225024 and No. 81871329, the Shanghai 2023 'Explorer Plan' No. 23TS1400400, the Pudong New Area Science and Technology Development Fund No. PKJ2023-Y53, and the Clinical Research Project of Shanghai Sixth People's Hospital No. ynhg202421. Funding agencies did not play a role in study design, data collection, analysis and interpretation, and manuscript writing.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Campbell BCV, Khatri P. Stroke. *Lancet*. 2020;396(10244):129–142. doi:10.1016/S0140-6736(20)31179-X
2. Feigin VL, Stark BA, Johnson CO, et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*. 2021;20(10):795–820. doi:10.1016/S1474-4422(21)00252-0
3. Tu WJ, Wang LD; on behalf of the Special Writing Group of China Stroke Surveillance Report. China stroke surveillance report 2021. *Military Med Res*. 2023;10(1):33. doi:10.1186/s40779-023-00463-x
4. Tu WJ, Zhao Z, Yin P, et al. Estimated burden of stroke in China in 2020. *JAMA Network Open*. 2023;6(3):e231455. doi:10.1001/jamanetworkopen.2023.1455
5. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2019;50(12). doi:10.1161/STR.0000000000000211

6. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J biomed informat*. 2018;77:34–49. doi:10.1016/j.jbi.2017.11.011
7. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol*. 2021;21(1):234. doi:10.1186/s12874-021-01416-5
8. Brooks N. How to undertake effective record-keeping and documentation. *Nurs Stand*. 2021;36(4):31–33. doi:10.7748/ns.2021.e11700
9. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt. Accessed March 16, 2024.
10. GPT-4. Available from: https://openai.com/gpt-4. Accessed February 16, 2024.
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
12. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology*. 2023;307(5):e231259. doi:10.1148/radiol.231259
13. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*. 2023;309(1):e231147. doi:10.1148/radiol.231147
14. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353–1365. doi:10.1227/neu.0000000000002632
15. Li D, Kao Y, Tsai S, et al. Comparing the performance of CHATGPT GPT -4, Bard, and Llama-2 in the Taiwan psychiatric licensing examination and in differential diagnosis with multi-center psychiatrists. *Psychiatry Clin Neurosci*. 2024:pcn.13656. doi:10.1111/pcn.13656
16. Our next-generation model: Gemini 1.5. Google. Available from: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/. Accessed February 16, 2024.
17. ZhiPuAI. Available from: https://zhipuai.cn/en/devday. Accessed February 16, 2024.
18. iFlytek starfire cognitive big Model-AI big language model-starfire big model-university of science and technology of China iFlytek. https://xinghuo.xfyun.cn/sparkapi. Accessed February 16, 2024.
19. Wu T, He S, Liu J, et al. A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA J Autom Sinica*. 2023;10(5):1122–1136. doi:10.1109/JAS.2023.123618
20. Cheng K, Li Z, Li C, et al. The potential of GPT-4 as an AI-powered virtual assistant for surgeons specialized in joint arthroplasty. *Ann Biomed Eng*. 2023;51(7):1366–1370. doi:10.1007/s10439-023-03207-z
21. Nishi H, Oishi N, Ishii A, et al. Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke*. 2019;50(9):2379–2388. doi:10.1161/STROKEAHA.119.025411
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845. doi:10.2307/2531595
23. Sun X, Xu W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*. 2014;21(11):1389–1393. doi:10.1109/LSP.2014.2337313
24. 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(3):e231362. doi:10.1148/radiol.231362
25. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. *Int J Med Inform*. 2024;183:105321. doi:10.1016/j.ijmedinf.2023.105321
26. Zaidat OO, Castonguay AC, Linfante I, et al. First pass effect: a new measure for stroke thrombectomy devices. *Stroke*. 2018;49(3):660–666. doi:10.1161/STROKEAHA.117.020315
27. Lehnen NC, Dorn F, Wiest IC, et al. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. Anzai Y, ed. *Radiology*. 2024;311(1):e232741. doi:10.1148/radiol.232741