Artificial Intelligence In Medicine 162 (2025) 103078

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought

Zhanzhong Gu [a,*], Wenjing Jia [a,*], Massimo Piccardi [a], Ping Yu [b]

[a] School of Electrical and Data Engineering, University of Technology Sydney, NSW, 2007, Australia
[b] School of Computing and Information Technology, University of Wollongong, NSW, 2522, Australia

ARTICLE INFO

Keywords:
Large language models
Automated clinical assessment
Generation-augmented retrieval
Hierarchical chain-of-thought
Electronic health record
Prompting strategy

ABSTRACT

Background: Understanding and extracting valuable information from electronic health records (EHRs) is important for improving healthcare delivery and health outcomes. Large language models (LLMs) have demonstrated significant proficiency in natural language understanding and processing, offering promises for automating the typically labor-intensive and time-consuming analytical tasks with EHRs. Despite the active application of LLMs in the healthcare setting, many foundation models lack real-world healthcare relevance. Applying LLMs to EHRs is still in its early stage. To advance this field, in this study, we pioneer a generation-augmented prompting paradigm "GAPrompt" to empower generic LLMs for automated clinical assessment, in particular, quantitative stroke severity assessment, using data extracted from EHRs.

Methods: The GAPrompt paradigm comprises five components: (i) prompt-driven selection of LLMs, (ii) generation-augmented construction of a knowledge base, (iii) summary-based generation-augmented retrieval (SGAR); (iv) inferencing with a hierarchical chain-of-thought (HCoT), and (v) ensembling of multiple generations.

Results: GAPrompt addresses the limitations of generic LLMs in clinical applications in a progressive manner. It efficiently evaluates the applicability of LLMs in specific tasks through LLM selection prompting, enhances their understanding of task-specific knowledge from the constructed knowledge base, improves the accuracy of knowledge and demonstration retrieval via SGAR, elevates LLM inference precision through HCoT, enhances generation robustness, and reduces hallucinations of LLM via ensembling. Experiment results demonstrate the capability of our method to empower LLMs to automatically assess EHRs and generate quantitative clinical assessment results.

Conclusion: Our study highlights the applicability of enhancing the capabilities of foundation LLMs in medical domain-specific tasks, *i.e.*, automated quantitative analysis of EHRs, addressing the challenges of labor-intensive and often manually conducted quantitative assessment of stroke in clinical practice and research. This approach offers a practical and accessible GAPrompt paradigm for researchers and industry practitioners seeking to leverage the power of LLMs in domain-specific applications. Its utility extends beyond the medical domain, applicable to a wide range of fields.

1. Introduction

Hospitals and medical practices around the world have increasingly adopted electronic health record (EHR) systems, resulting in massive amounts of electronic patient data in both structured (*e.g.*, disease codes, medication codes) and unstructured (*i.e.*, clinical narratives such as progress notes) formats. The advancements in AI techniques, including machine learning, deep learning, and natural language processing (NLP), have provided researchers with powerful techniques to automate the methods and process of secondary data analysis to support clinical decisions and research based on these massive amounts of EHR data [1–4]. Currently, the EHR data analytic methods encounter several significant limitations. These include the requirement for large volumes of labeled datasets for model training, the necessity for entity (health terms) and relationship annotation, labor-intensive preprocessing procedures, and inadequate quantitative assessment capabilities [1,3,5].

The recent large language models (LLMs) hold remarkable capability in natural language understanding (NLU) and natural language inference (NLI) [6,7]. They can comprehend and answer questions directly for a given text, surpassing the classical machine learning and deep learning methods, which require sentence-by-sentence or word-by-word processing and annotation [8]. Therefore, these LLMs are

* Corresponding authors.
E-mail addresses: Zhanzhong.Gu@uts.edu.au (Z. Gu), Wenjing.Jia@uts.edu.au (W. Jia).

https://doi.org/10.1016/j.artmed.2025.103078
Received 4 April 2024; Received in revised form 18 January 2025; Accepted 3 February 2025
Available online 12 February 2025

highly promising AI techniques for enhancing EHR analytic technologies to improve the quality and productivity of healthcare services. However, it remains a challenge to directly apply these LLMs in real-world domain-specific tasks [9,10], because most generic LLMs are trained on general language data and lack domain-specific knowledge [11], while the very few medical domain LLMs are proprietary and not publicly available [12–14]. Also, there is little report about the application of LLMs in quantitative clinical assessment tasks.

Previous studies have demonstrated that with appropriately designed prompting strategies, generic LLMs can achieve comparable performance to domain-specific LLMs without the time-consuming and costly training or fine-tuning of LLMs [15–17]. Therefore, we explored the feasibility of applying prompting techniques to enable generic LLMs to complete our clinical assessment task of stroke severity. However, our initial research has found that there are several main challenges of applying generic LLMs directly for automated stroke assessment. These include the evaluation of the applicability of foundation LLMs, the lack of stroke assessment knowledge, the limited context length in processing large EHRs, the inaccuracy of reasoning quantitative assessment results, and the inevitable hallucination during the generation. By leveraging the in-context learning (ICL) ability of LLMs, in this paper, a series of prompting strategies, including prompt-driven LLM selection, generation-augmented knowledge base construction, summary-based generation-augmented retrieval (SGAR), hierarchical chain-of-thought (HCoT), and an ensembling mechanism, are developed to tackle these issues, empowering LLMs for automated quantitative clinical assessment from EHRs (see Fig. 1).

First, with the popularity of LLMs, a plethora of new models are continuously emerging. However, their applicability and performance need to be carefully evaluated in quantitative clinical assessment tasks. Thus, in this paper, first and foremost, an effective and efficient prompting-based LLM selection approach is developed. Next, to enhance the LLM's knowledge of stroke assessment, generation-augmented retrieval (GAR) is a suitable solution [18,19]. This process first constructs an external knowledge base compromising stroke assessment guidelines and demonstrations, using the well-established National Institutes of Health Stroke Scale (NIHSS) [20] as the quantitative stroke assessment standard, and generating demonstrations from expert-validated assessment results on a labeled EHR dataset CSCR [21]. We further develop an innovative summary-based GAR (SGAR) method to enhance the retrieval of corresponding assessment criteria and demonstrations, thereby facilitating LLMs' generation process and reasoning performance. Subsequently, a novel HCoT prompting strategy that integrates the document-level macro sequential chain [22] and sentence-level micro CoT [23], is proposed to overcome the challenges of LLM's limited context length in processing large EHRs, and improve the performance of LLM inference. Using the popular Langchain library [22], large EHRs are split into short sentences and sequentially processed by the macro chain. Meanwhile, the micro CoT is capable of significantly improving the performance of LLM inference through logical steps provided in the demonstrations. Finally, an ensembling strategy is applied to integrate multiple generation results to control the impact of LLM's hallucination in generation.

We highlight the following contributions of our research:

(1) A GAPrompt Paradigm for Clinical Assessment: We develop an overarching generation-augmented prompting paradigm (GAPrompt), which effectively extends the capabilities of generic LLMs, empowering them in clinical domain-specific applications, enabling automated and quantitative assessment of stroke severity with enhanced accuracy and efficiency.

(2) A Prompt-Driven LLM Selection Method: We propose an effective and efficient prompting-based LLM selection process to evaluate and identify the most suitable LLM for quantitative clinical assessment tasks.

(3) A Summary-based Generation-Augmented Retrieval (SGAR) Method: We develop an SGAR method based on LLM-generated summaries of assessment guidelines, demonstrations, and query EHRs to improve performance.

(4) A Hierarchical Chain-of-Thought (HCoT) Prompting Strategy: We propose a novel HCoT prompting strategy to address the limitations of the context length of LLMs by integrating a macro sequential chain at the document level with a micro-coT at the sentence level, breaking down the task step by step and improving reasoning performance.

(5) An Ensembling Strategy for Enhanced Robustness: We apply an ensembling strategy that integrates multiple generation results to mitigate the impact of LLM hallucinations, enhancing the reliability of LLM-generated outputs by ensembling diverse inferences into a robust final result.

The remainder of this paper is organized as follows: Section 2 provides a concise overview of existing LLM, GAR and HCoT prompting strategies. Section 3 details our methods of the GAPrompt paradigm, including the prompt-driven LLM selection, generation-augmented knowledge base construction, summary-based generation-augmented retrieval, hierarchical chain-of-thought, and the ensembling approach. The experiment design and results are presented in Sections 4 and 5, respectively, followed by a discussion and conclusion in Section 7.

2. Related works

The techniques used in this research include state-of-the-art LLMs, cutting-edge techniques on generation-augmented retrieval and prompting strategies, especially the hierarchical chain-of-thought and ensembling.

2.1. Large language models (LLMs)

Large language models (LLMs) refer to the foundation language models that can understand and generate natural language. They are based on the transformer architecture [24] and pre-trained on a large amount of data, typically containing hundreds or billions of parameters [25]. These include GPT-3.5 [7], GPT-4 [26], Meta's Llama model [27], Google's PaLM model [13], etc.

Many advanced proprietary LLMs have exhibited versatility in handling a wide array of tasks, including those in the field of health and medicine [28–30]. Furthermore, specific LLMs have been meticulously fine-tuned for medical applications, such as Med-PaLM [13], and Med-PaLM 2 [14]. This dual capability of general applicability and domain-specific refinement underscores the potential of LLMs in health and medicine.

Currently, some open-source LLMs have demonstrated excellent performance even comparable to state-of-the-art (SOTA) proprietary LLMs across various tasks [31,32]. These models include LlaMa2 [27], BLOOM [33], Falcon [34], Alpaca [35], MedAlpaca [36], and many notable open-source Chinese LLMs, such as Baichuan [37], Qwen [38] and XVERSE [39]. Some LLMs with fewer parameters are specifically fine-tuned on Chinese medical data, such as DoctorGLM [40] and HuatuoGPT [41].

Performance assessments of these models are typically conducted on datasets of specific tasks, such as MMLU [42], MBPP [43], GSM8K [44], and Math [45], to test the model's multilingual knowledge capabilities, translation, mathematical reasoning, coding, and other capabilities [46,47]. However, these evaluations may not be adequate for identifying the applicability and performance of LLMs in real-world applications such as clinical assessment using EHRs. To address this, we designed a set of prompt-driven LLM selection templates to effectively identify a foundation LLM that aligns with specific task requirements (see Section 3.1).

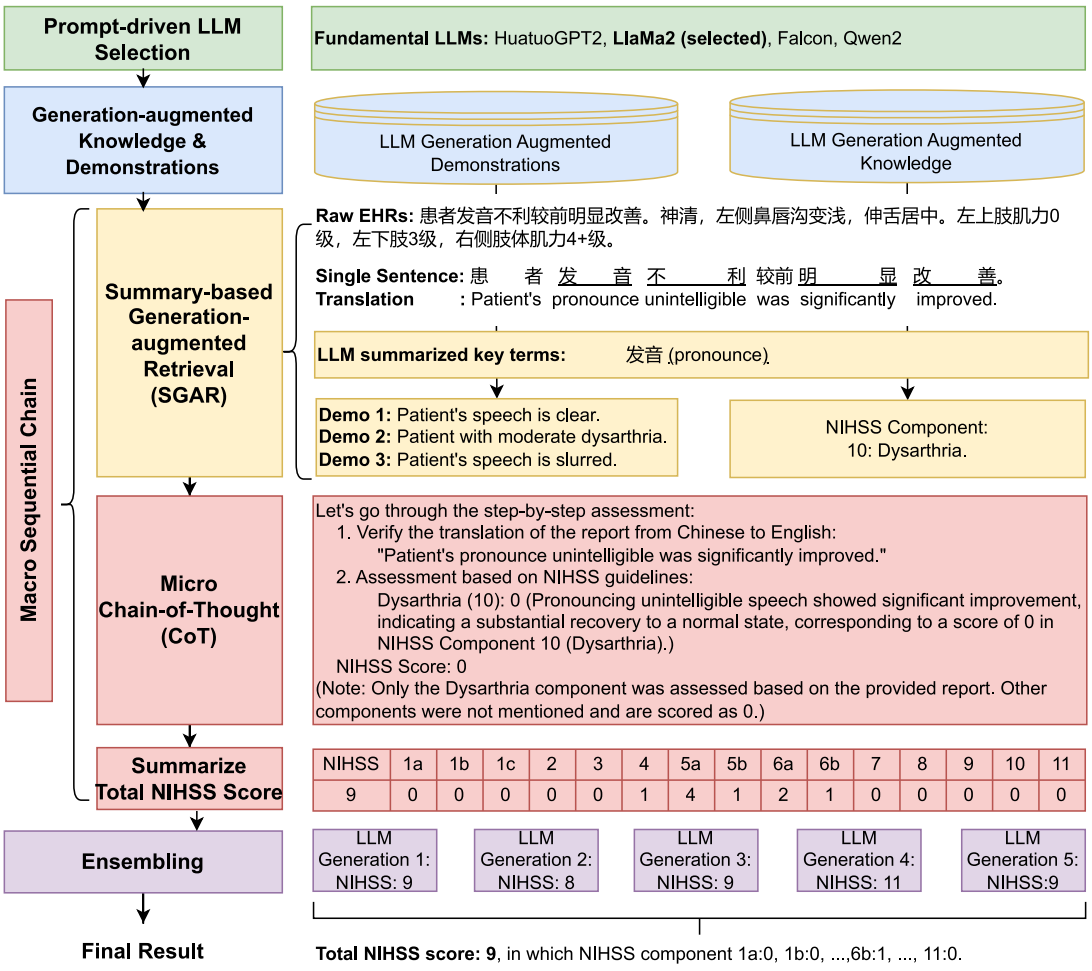

Fig. 1. The architecture of our proposed GAPrompt paradigm. Green color: prompt-driven LLM selection; blue color: generation-augmented knowledge base construction; orange color: summary-based generation-augmented retrieval (SGAR); red color: hierarchical chain-of-thought (HCoT); purple color: ensembling of multiple generations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Generation-augmented retrieval (GAR)

Retrieval is a technique aimed at enhancing the performance of LLM generation by retrieving valuable information and demonstrations from an external knowledge base. This external knowledge base can be existing databases or structured resources with domain-specific knowledge [48,49]. However, building and maintaining a knowledge base suitable for LLMs is labor-intensive and demands significant human and time resources. This effort is also susceptible to errors and omissions, subsequently impacting the effectiveness of the generated content in various tasks [50]. Leveraging the powerful generation capability of LLMs, the technique of generation-augmented construction of the external knowledge base has been proposed to address the above challenges and proved to be very effective [30,51].

Various retrieval methods can be employed to extract content relevant to the query from the knowledge base. These include classic statistical matching methods such as BM25 [52], and embedding-driven retrieval mechanisms, like KNN [30] and dense representation-based retrieval (DPR) [53]. However, our preliminary research indicates that these retrieval methods often treat each word in the query equally, resulting in failure to precisely identify records most relevant to the "keywords". Generation-augmented retrieval (GAR) is thus introduced to mitigate these limitations by enhancing the semantics of queries, leading to a substantial improvement in retrieval accuracy [19].

This study employs a generation-augmented approach to construct an external knowledge base for stroke assessment. By leveraging the generative capabilities of LLMs and referencing dataset labels, our approach ensures the efficient generation of a high-quality external knowledge base. Furthermore, to ensure a high retrieval accuracy, we develop an innovative summary-based GAR method to replace the full-text embeddings with LLM-generated summary indexes that extract and embed only the critical terms. This effectively enhances retrieval accuracy and the overall task performance.

2.3. Prompting strategies

Prompt engineering entails the strategic design of effective prompts to guide LLMs in accomplishing downstream tasks. It plays a pivotal role in successful LLM generation. Existing prompting strategies include few-shot learning, chain-of-though (CoT), and ensembling methods.

Few-shot learning is a key in-context learning (ICL) capability of LLMs [8]. It teaches an LLM to learn from only a small number of labeled examples to generate a new, unseen but similar result. Chain-of-thought learning encourages an LLM to "think step by step", entering a mode of reasoning where it systematically breaks down complex tasks into a sequence of ordered steps. This prompting method has improved the accuracy and coherence of the generated output [23,50], and is entitled CoT to provide a vivid portrayal of the model's sequential thinking process. As for ensembling, it combines the outputs of multiple individual models or multiple generations by one model with different degrees of randomness to produce a more accurate and reliable result, instead of relying on a single reasoning output [30,54]. Well-designed

prompting strategies have demonstrated comparable or even superior performance than specific fine-tuning methods [30,55]. However, to date, there is little report on the successful implementation of the emerging prompt strategies in clinical assessment tasks using the EHR data.

To address the methodology gap, in this study we have developed a set of generation-augmented prompting strategies and formulated a prompting paradigm entitled "GAPrompt". This paradigm is designed to empower generic foundational LLMs to quantitatively assess clinical EHRs.

3. Methods

We propose GAPrompt that progressively enhances the capabilities of the generic LLMs for our stroke assessment application. It specifically addresses the limitations of LLMs, including their uncertain applicability, lack of stroke assessment knowledge, limited context length, inaccuracy in quantitative reasoning, and the issue of hallucination.

As shown in Fig. 1, our proposed GAPrompt paradigm comprises five process components: (i) prompt-driven LLM selection (in green), (ii) generation-augmented knowledge and demonstration construction (in blue), (iii) SGAR (in orange), (iv) HCoT (in red), and (v) ensembling of multiple generations (in purple). The right part of Fig. 1 illustrates the automated NIHSS scoring process using GAPrompt with specific examples and step-by-step reasoning.

3.1. Prompt-driven LLM selection

To evaluate the applicability of candidate generic LLMs for our specific application scenario, we first devise a prompt-driven LLM selection strategy (see Fig. 1). In this strategy, we create six prompt templates to evaluate the capabilities of candidate LLMs in the following six aspects: the foundational knowledge required for stroke assessment ("Knowledge"), comprehension of stroke-related knowledge and memory capacity ("Comprehension"), learning from the few-shot examples about stroke ("Learning"), chain-of-thought (CoT) reasoning ("Reasoning"), ensuring consistency in the generated outputs ("Consistency"), and controlling hallucinations ("Anti-hallucination"). Fig. 2 presents examples of the detailed format of each prompt template.

In these examples, the **Knowledge prompt**, "Tell me the definition of the National Institute of Health Stroke Scale (NIHSS) and its scoring criteria", requires a highly specialized response. It assesses an LLM's foundational knowledge in stroke assessment using NIHSS. In the **Comprehension prompt**, we first present a comprehensive definition of NIHSS along with its scoring criteria, afterwards we pose a similar question to evaluate the LLM's comprehension based on the given context. The **Learning prompt** presents examples in a question–answer format and concludes with a similar question to check if the LLM can learn from these examples. In the **Reasoning prompt**, we provide a logical reasoning demonstration in question-answer form, followed by a similar question to assess the LLM's capability to learn logical reasoning from examples. The **Consistency prompt** repeats a question using different expressions to examine the consistency of the LLM's responses. Finally, in the **Anti-hallucination prompt**, we pose an initial question and then ask an unrelated one (*e.g.*, "The patient's speech is unclear. So, what is the patient's muscle strength level on the left leg?") to evaluate the LLM's hallucination control ability.

While these prompts may not comprehensively assess an LLM's capabilities, they establish a fast and systematic method to evaluate the performance of generic LLMs in the specific context of stroke severity assessment and identify the foundation LLM that meets our task requirements. To quantitatively compare the selection results, we utilize the widely adopted Exact Match (EM) method and employ the EM score as the evaluation metric [53]. Detailed explanations of this approach are provided in Sections 4.1.1 and 4.2. With this process, we have identified the most suitable model from four candidate LLMs (see Section 5.1 for details).

3.2. Generation-augmented knowledge base construction

Two types of external knowledge are required for LLMs to effectively perform the task of quantitative assessment of stroke using EHRs: task-specific knowledge and highly relevant demonstrations. The former refers to the measurable NIHSS assessment criteria, and the latter are the examples given to the LLMs for task execution.

Task-specific Knowledge. In our evaluation of LLM performance during the prompt-driven LLM selection process (Section 3.1), we have observed that, while LLMs possess a fundamental understanding of stroke assessment, they struggle with consistently identifying assessment items and assigning precise NIHSS scores in reasoning. Therefore, we integrate an explicit NIHSS assessment guideline[1] as an external task-specific knowledge to improve the performance of the foundation LLM in this task. The NIHSS assessment protocol comprises 11 components, each with distinct assessment objectives and scoring criteria.

LLM-generated Demonstrations. Previous research works have featured the significance of using demonstrations to improve the performance of LLMs in text generation tasks [14,23]. They have also explored the potential of substituting manually composed examples with LLM-generated demonstrations. In accordance with the findings that LLMs can automatically generate CoT examples and make corrections based on the given ground truth [15,30,50], we introduce the following prompt template, as shown in Fig. 3, for LLMs to generate demonstrations.

3.3. Summary-based generation-augmented retrieval (SGAR)

Retrieval is a pivotal step in our prompting approach. Previous research has shown that dynamic retrieval, which takes into account the content of each query to accurately retrieve highly relevant demonstrations, significantly improves the overall quality of CoT prompting [30,50]. Furthermore, the GAR method [19] that uses LLMs to augment the query content has proven effective in enhancing retrieval accuracy. In light of these insights, we propose an innovative summary-based GAR (SGAR) approach that employs LLM-generated summaries to improve the retrieval accuracy.

Unlike previous methods that focused solely on enhancing the input query, our approach introduces the concurrent LLM-generated summarization of both the input query and the external knowledge base. This dual summarization approach enables the query to capture essential information in the sentence-level queries, and compress the information at the document or paragraph level in the knowledge base.

The detailed process of SGAR includes the following steps: (1) **Define Summarization Criteria:** We establish summarization criteria that prioritize extracting information related to anatomy, inspection, and symptoms, focusing on key terms relevant to stroke severity assessment. (2) **Generate Summaries with LLMs:** Using the defined criteria, we instruct the LLM to generate summaries for the knowledge, demonstrations and the EHR. For knowledge and EHR records, we apply LLM-based summarization to each record. For demonstrations, we summarize only the question portion to enable effective query matching. (3) **Embed Summaries as Metadata:** After summarization, both the knowledge summaries and demonstration summaries are embedded using sentence-transformers [56]. These embeddings are saved as metadata in the knowledge base, corresponding to their respective records. (4) **Retrieve and Match Metadata:** During the retrieval process, the algorithm searches for and matches the summarized query with the metadata of the knowledge and demonstrations in the database. (5) **Return Demonstrations and Knowledge:** Upon successful matching, the algorithm retrieves and returns the raw knowledge and demonstrations, ensuring that no information is lost from the knowledge base during this process.

[1] https://www.ninds.nih.gov/health-information/public-education/know-stroke/health-professionals/nih-stroke-scale

Knowledge

Instruction: Tell me the definition of the National Institutes of Health Stroke Scale (NIHSS) and its assessment criteria.
Input: None.

Comprehension

Instruction: Tell me the definition of NIHSS and its assessment criteria based on the given information.
Input: **{{assessment criteria}}**

Learning

Question: Which NIHSS component is for the assessment of Dysarthria?
Answer: The 10th component of NIHSS.
Question: What does the 10th component of NIHSS assess?

Reasoning

Question: Muscle strength levels 1 to 5 score 4 to 0 in NIHSS, respectively. What does level 3 score?
Answer: **Let's think step by step.** Level 3 is the 3rd level, thus it scores the third value in the range of 4 to 0, which is 2.
Question: what is the Level 1's score?

Consistency

Question: NIHSS has 11 assessment components. What is the 11th component?
Question: What is the last component of NIHSS?

Anti-hallucination

Question: Tom has unclear speech. What is his limb muscle strength level?
Answer: []

Fig. 2. The six prompt templates applied to select the optimal foundation LLM. Six capabilities of LLMs, including Knowledge, Comprehension, Learning, Reasoning, Consistency, and Anti-hallucination, are evaluated using these defined prompts.

Prompt Template for LLMs to Generate Demonstrations

Context:**{{Knowledge (assessment criteria)}}**
Instruction: Please follow the assessment criteria to assess the scores of each NIHSS component from the following report.
{{EHR sentence}}
Let's think step by step.
1. If the report is not in English, translate it to English first.
2. Determine which components of NIHSS are related to the report, and assess the score.
3. Not mentioned components score 0.
4. Correct the answer according to the ground truth for each component:
{{Ground truth}}

Fig. 3. The template for LLMs to generate demonstrations.

Micro Chain-of-Thought Template

Context:**{{Knoledge (assessment criteria)}}**
Demo: **{{Demostrations}}**
Input: **{{EHR sentence}}**
Instruction: Please follow the assessment criteria to assess the scores of each NIHSS component from the given report.
Let's think step by step.
1. If the report is not in English, translate it to English first.
2. Determine which components of NIHSS are related to the report, and assess the score.
3. Not mentioned components score 0.

Fig. 4. An example of the micro CoT prompting template.

3.4. Hierarchical chain-of-thought

To address the limitations of the foundation LLMs that we have encountered, including limited context length and inference errors, we introduce two techniques — macro sequential chain and micro CoT, and encapsulate them in our method entitled Hierarchical Chain-of-Thought (HCoT) (see Fig. 1). Chain-of-thought prompting has been empirically validated as an effective method for prompting LLMs. It enables systematic reasoning in alignment with the logic flow of the few-shot examples [23,50].

In our proposed HCoT mechanism, the macro sequential chain is employed to decompose the complex assessment tasks into a series of sequential steps, breaking down EHRs at the page or paragraph level into manageable sentence-level reasoning. The micro CoT then performs sentence-level generation, effectively overcoming the context length limitation while improving the accuracy of the responses (see the red components in Fig. 1).

3.4.1. Macro sequential chain

Leveraging the Langchain platform [22], we traverse each EHR data through four sequential chains, *i.e.*, splitting, translation, retrieval, and micro CoT (see Fig. 4). The output of one chain serves as the input for the next chain. Distinct prompt templates are applied at different chains to achieve each one's intended purpose.

The EHR dataset used in this study, *i.e.*, the CSCR dataset [21], is provided by a hospital in China, thus in Chinese language. First, the splitter splits the paragraphs in the EHR dataset into short sentences. Then, the translator translates each sentence into English. During the retrieval process, each input sentence is processed by the foundation LLM to first summarize the content (as described in Section 3.3). Then, the compressed content is fed into the retriever to retrieve the relevant contextual knowledge and demonstrations from the external knowledge base. Finally, the raw sentences, the retrieved knowledge and demonstrations are all fed into the next chain for micro CoT learning.

3.4.2. Micro Chain-of-Thought (CoT)

Micro CoT is the core step of our proposed GAPrompt prompting paradigm (see Fig. 1 in red). Fig. 4 provides a detailed illustration of the micro CoT prompting template. Unlike the existing CoT methods that use a fixed set of examples for few-shot prompting [13,23], our sentence-level micro CoT is underpinned by an external knowledge base in addition to the demonstrations. We first incorporate the standard NIHSS assessment criteria into the prompt template as contextual knowledge, addressing inconsistencies caused by LLM's potential uncertainty and hallucination. Then, given the limited context length of the foundation LLM, we employ a three-shot prompting approach, restricting the number of demonstrations to three instances. Furthermore, our inference process aligns with the CoT logic shown in the demonstrations. Experiment results show the effectiveness of CoT to improve LLMs' EHR analyzing performance.

Table 1
The performance of candidate LLMs with six prompting templates, using the evaluation metric of exact match (EM).

LLMs	Knowledge	Comprehension	Learning	Reasoning	Consistency	Anti-hallucination	Overall
LlaMa2-70B [27]	0.48	0.73	0.46	0.71	0.89	0.67	**0.66**
Qwen-72B [38]	0.37	0.48	0.37	0.65	0.90	0.66	0.57
Falcon-40B [34]	0.38	0.43	0.35	0.60	0.85	0.70	0.47
HuatuoGPT2-34B [41]	0.40	0.36	0.28	0.56	0.80	0.57	0.50

3.5. Ensembling

We implement an ensembling strategy inspired by prior research [31]. This approach involves varying the LLM's temperature parameter to generate diverse outputs across multiple inference runs. Specifically, for each input EHR dataset, the LLM is independently prompted five times, producing multiple outputs. These outputs are then aggregated through a majority voting process, ensuring robustness and reducing the likelihood of errors in the final result.

In summary, the GAPrompt paradigm starts with identifying the most suitable LLM for the task and constructing an external knowledge base with LLM augmentation. Afterward, the paradigm retrieves the task-specific knowledge, *i.e.*, NIHSS assessment criteria, and demonstrations according to the input EHR report. Based on the retrieval outputs, the LLM inference is conducted using HCoT prompting and an ensembling strategy. The assignment of the NIHSS score for each component is finally carried out.

4. Experiment design

In this section, we first introduce the experiment datasets, which include the samples for LLM selection, the generation-augmented knowledge base, and the test dataset for stroke severity assessment, followed by evaluation metrics.

4.1. Datasets

4.1.1. Samples for selecting the foundation LLM

As detailed in Section 3.1, we have designed six task-specific prompt templates to evaluate the capabilities of LLMs in our stroke assessment use case. Each prompt template contains a specified query paired with a corresponding ground truth answer. The candidate LLMs are individually loaded and presented with each prompt in sequence. Their performance is assessed by comparing the outputs to the respective ground-truth answers using the EM score [53].

4.1.2. Generation-augmented knowledge base

We construct a knowledge base composing both task-specific knowledge and demonstrations (see Section 3.2, with its detailed distribution of the knowledge base for each NIHSS component shown in Appendices Table A.1). The knowledge is referred to as the detailed definitions and scoring criteria for the 11 NIHSS components, which are further decomposed into 15 sub-components.

The LLM-generated demonstrations utilize the original EHR data from the CSCR datasets [21], which contains EHRs from 1931 patients. These EHRs are split into sentences, each with expert-validated NIHSS scores. After removing duplicate sentences and unrelated items, we are left with 3314 sentence-level demonstrations generated by the LLM prompt templates illustrated in Fig. 3.

We employ a commonly used sentence-transformer embedding, "all-mpnet-base-v2" [56], to convert both the assessment criteria of each NIHSS component and the demonstrations into sentence vectors and then store these vectors as a knowledge base, where the LLM-generated summaries are stored as metadata indexes for subsequent retrieval.

4.1.3. Test dataset for stroke severity assessment

Our test dataset [21] comprises ground-truth stroke assessment scores for 33 patients, including both macro and micro-level ground truth. Table A.2 shows the detailed distribution of the test dataset for all NIHSS components. The micro-level samples represent the sentence-level inferencing result generated by LLMs with the micro CoT. The macro-level ground truth refers to the patient-level assessment scores from the given 33 EHRs. The macro chain summarizes the micro sentence-level results of the same patient.

4.2. Evaluation metrics

Following prior research [53], we utilize the Exact Match (EM) score to assess the performance of the LLM selection and evaluate the retrieval performance using Top-k retrieval accuracy. To evaluate the performance of quantitative stroke assessment, we adopt the widely-used F1 score metric [57].

EM score is the proportion of the predicted answer texts that are identical to the ground-truth answer, after string normalization such as article and punctuation removal.

Top-k Retrieval Accuracy is defined as the proportion of questions for which the top-k retrieved records contain at least one correct answer. This metric sets up the upper bound of how many relevant questions are extracted by the retriever.

F1 score is the harmonic mean of precision and recall, offering a balanced assessment of a model's performance. Precision measures the proportion of true positive predictions made by the model. Recall, on the other hand, quantifies the proportion of actual positive instances that the model identified. By combining precision and recall, the F1 score considers both false positives and false negatives, providing a comprehensive measure of the model's performance.

5. Results

5.1. Evaluation of LLM selection

In Section 3.1, we have devised six prompting templates for selecting a candidate pool of LLMs. Table 1 shows the results of LLM selection using the EM score (see Section 4.2).

From Table 1, we can see that LlaMa2-70B and Qwen-72B exhibit superior overall abilities compared to their competitors. Notably, they demonstrate a strong ICL ability for learning knowledge from the external knowledge base and understanding the logic from the demonstrations. In contrast, the specifically fine-tuned medical domain LLM HuatuoGPT2 does not perform well, especially in the learning ability. This may be attributed to the nature of our task, which goes beyond a simple medical Q&A task but fully utilizes ICL and CoT to comprehend in-context knowledge and infer from the retrieved knowledge and demonstrations. This test result leads to our selection of the most powerful foundation LLM, *i.e.*, LlaMa2-70B, for our task.

5.2. Evaluation of SGAR

Table 2 shows the comparison results of adopting different retrieval methods to retrieve both task-specific knowledge and the demonstrations from our constructed knowledge base (see Table A.1), using Top-k retrieval accuracy as the evaluation metric. Two retrieval methods, the statistical BM25 [52] and vector database as retriever (Vdb) [22], are evaluated and compared with the adoption of SGAR.

The experiment results reveal that the Vdb method outperforms the traditional BM25 with higher retrieval accuracy on all top-k settings.

Table 2
Top-k retrieval accuracy (%) on both task-specific knowledge and the demonstrations.

Method	Knowledge			Demonstrations		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
BM25 [52]	6.76	38.12	49.23	35.26	47.88	54.52
BM25+SGAR	22.83	43.88	53.13	48.76	54.64	61.74
Vdb [22]	60.71	73.92	78.61	71.46	82.64	88.12
Vdb+SGAR	67.48	76.34	**80.25**	79.84	87.81	**93.51**

When applying SGAR to both retrieval methods, the summarized key terms including anatomy, inspection and symptoms, are extracted from the raw EHR as queries, thus yielding more accurate retrieval outcomes compared with using the whole EHR as queries. On the other side, SGAR significantly reduces the context length occupied by retrieval and LLM inferencing. Compared to retrieving the entire knowledge base, the summarized indexes save a large portion of context occupancy, highlighting the efficiency and effectiveness of SGAR.

5.3. Evaluation of micro cot

Table 3 shows the F1 scores of LLM inferencing with micro CoT on the micro-level testing dataset. The temperature parameter is set to 0.01 for all four LLMs for fair comparison.

Four foundation LLMs, including LlaMa2-70B [27], Qwen-72B [38], Falcon-40B [34] and HuatuoGPT2-34B [41], are tested to quantitatively assess EHR sentences and generate the scores for each NIHSS component. All models demonstrate excellent accuracy of above 80% in assigning NIHSS scores for each assessment component based on sentence-level EHRs, demonstrating the effectiveness of micro CoT in enhancing the performance of generic LLMs in our clinical domain-specific tasks. Among these LLMs, LlaMa2-70B shows the best performance with a 94.64% F1 score, slightly surpassing Qwen-72B and largely superior to the Falcon-40B and HuatuoGPT2-34B. The result is consistent with the LLM selection results presented in Section 5.1, reaffirming the effectiveness of our fast but efficient prompt-driven LLM selection approach.

A comprehensive error analysis revealed that most errors occurred when NIHSS assessment components that should have been assigned a score of "Not mentioned" (scored 0) were instead assigned a non-zero value by the LLM generation. This issue arises due to the LLM's tendency to infer scores for multiple components based on symptoms relevant to just one. Although our micro CoT process manages to control this through detailed instructions and demonstrations, it still cannot completely eliminate the inherent randomness and hallucination of the LLMs. To mitigate this issue, leveraging more advanced LLMs, refining prompts, and incorporating post-processing strategies such as ensembling may serve as effective solutions.

Table 3
Comparison of the F1 scores (%) obtained from different foundation LLMs with micro CoT (temperature = 0.01).

NIHSS component	LlaMa2-70B [27]	Qwen-72B [38]	Falcon-40B [34]	HuatuoGPT2-34B [41]
1a	94.21	94.63	84.62	88.16
1b	95.87	93.60	85.14	84.33
1c	98.69	98.34	85.66	85.00
2	95.05	94.49	81.22	80.88
3	96.38	95.57	83.34	80.43
4	91.83	94.18	79.25	80.61
5a	96.57	96.70	86.15	84.25
5b	94.73	94.44	77.12	81.54
6a	97.17	97.28	80.53	78.27
6b	95.18	94.71	77.84	77.52
7	86.05	86.56	70.53	72.28
8	86.14	86.24	70.11	74.63
9	94.84	93.84	78.62	82.22
10	97.95	97.67	84.51	82.79
11	98.99	98.90	85.12	83.93
Overall	**94.64**	94.48	80.65	81.12

Table 4
Comparison of the F1 scores (%) of macro sequential chain obtained with different foundation LLMs (0.01 temperature).

NIHSS component	LlaMa2-70B [27]	Qwen-72B [38]	Falcon-40B [34]	HuatuoGPT2-34B [41]
1a	78.37	70.44	63.12	65.28
1b	75.88	54.36	62.56	58.55
1c	89.47	87.88	60.44	61.23
2	90.71	90.71	67.20	72.65
3	95.24	98.46	71.33	71.44
4	62.07	82.44	72.05	74.64
5a	85.27	89.22	78.60	84.22
5b	81.77	91.02	77.97	91.24
6a	80.67	85.07	70.23	85.72
6b	76.16	83.24	71.54	81.84
7	77.44	85.72	81.35	80.88
8	69.03	63.30	50.51	69.63
9	81.82	56.57	65.54	68.75
10	97.11	94.43	87.88	92.51
11	98.46	96.88	66.40	70.35
Overall	**82.56**	81.98	69.78	75.26

5.4. Evaluation of macro sequential chain

In this section, we evaluate the F1 score of the macro sequential chain using the macro-level testing dataset (see Table A.2). The macro chain consists of five steps, including splitting, translation, retrieval, micro CoT and summarization (see Section 3.4.1). Utilizing micro CoT for LLM inference, the micro-level results associated with the same patient are aggregated to produce the macro-level outcome.

Table 4 shows the results of the macro sequential chain obtained on the basis of the LLM inferencing using micro CoT with four foundation LLMs. It indicates that macro results closely align with the micro CoT results, albeit slightly lower, achieving an F1 score of 82.56% for the best-performing LlaMa2-70B model at a temperature of 0.01. This occurs because, as revealed by the error analysis in Section 5.3, incorrect non-zero values may appear in the micro-level results. These non-zero errors may override the correct zero-value results through macro-level aggregation, resulting in a cumulative effect of errors that ultimately reduces the patient-level assessment accuracy. To mitigate this issue, we implemented an ensembling strategy (see Section 5.5), which has demonstrated promising improvements.

5.5. Final results with ensembling

Finally, after obtaining the initial assessment results with the prompting strategies including the SGAR and HCoT, the ensembling strategy is conducted through a voting process that selects the most common score among multiple LLM generations. We use the best-performing LlaMa2_70B as the foundation model and set five different temperature values for the generation. Table 5 shows the final result, an F1 score of 84.78%, of the quantitative stroke assessment after ensembling the five independent generations.

From Table 5, we can witness an overall 2.91% improvement in ensembling compared with the mean F1 score 81.87% of the five individual generations. The dramatically changing standard deviations from 0.77% to 10.15% illustrate severe fluctuations in the LLM generations under different temperature values, which is in line with the inevitable randomness and hallucination in LLM generations. Ensembling has proven to be highly effective in addressing issues where there

Table 5
The final result of the stroke assessment in NIHSS after ensembling (F1 score with %). Five independent generations are conducted based on LlaMa2-70B with different temperatures.

NIHSS component	Mean-F1 score	Standard deviation	Ensembled F1 score	Improvement
1a	75.07	3.24	78.37	3.30
1b	66.59	9.38	77.78	11.19
1c	89.14	1.17	89.47	0.33
2	87.90	2.88	90.71	2.81
3	95.11	3.77	95.24	0.13
4	70.93	7.25	77.56	6.63
5a	83.74	3.59	85.27	1.53
5b	82.27	6.10	86.53	4.26
6a	82.42	1.77	82.39	−0.03
6b	75.55	4.23	73.84	−1.71
7	81.17	4.79	77.44	−3.73
8	69.35	6.14	73.51	4.16
9	75.98	10.15	85.19	9.21
10	95.41	3.93	100.00	4.59
11	97.51	0.77	98.46	0.95
Overall	81.87	1.25	**84.78**	2.91

are significant differences in the results of multiple LLM generations. For instance, NIHSS components 1b and 9, which exhibit large standard deviations of 9.38% and 10.15%, respectively, show significant improvements of 11.19% and 9.21%, respectively after ensembling. For certain NIHSS components such as 10, the foundation LLM has achieved high F1 scores of 95.41%. After ensembling, the performance has reached an F1 score of 100%.

5.6. The effectiveness of individual components

Fig. 5 illustrates the effectiveness of each GAPrompt component, based on the overall F1 score of the quantitative assessment using LlaMa-70B on the test dataset.

Our proposed GAPrompt paradigm consists of generation-augmented knowledge base construction (represented as "Knowledge" in Fig. 5), the SGAR method to retrieve the knowledge and the demonstrations, the HCoT strategy that integrates micro CoT with macro sequential chain, and the ensembling strategy to incorporate the inference results from five generations. From Fig. 5, we can find how much each component of GAPrompt contributes to the overall results.

The blue bar shows the inferencing performance of the foundation LLM, with a moderate F1 score of 56.84%. It is not surprising since the inferencing largely relies on the basic knowledge of the foundation LLM, which is not entirely accurate as described in Section 3.1.

The green bar represents the performance improvement (+7.10% F1 score) achieved by importing task-specific knowledge during LLM inferencing. Two factors lead to this enhancement. The first is the inclusion of domain-specific knowledge, *i.e.*, the detailed NIHSS assessment criteria in this study, which clearly defines the assessment components of NIHSS. The second is the provision of the detailed scoring criteria, which enables the LLM to better understand the given EHR and conduct quantitative assessment.

The orange bar indicates the improvement (+3.85% F1 score) when employing our proposed GAR method based on the LLM-generated summary index. This improvement is compared with the performance using the full-text-based retrieval method. It is consistent with the findings in Section 5.2, indicating higher retrieval accuracy of our proposed SGAR on both knowledge and demonstration retrieval.

The most significant improvement of our GAPrompt falls on the HCoT strategy (+14.81% F1 score). It indicates that our designed promptings on both the macro sequential chain and the micro CoT contribute the most to empowering the foundation LLM in completing our task. This finding is consistent with the previous works, *i.e.*,

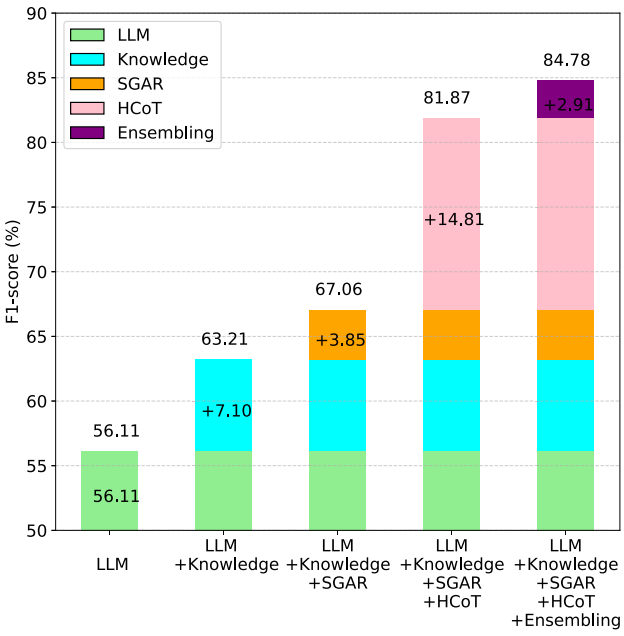

Fig. 5. The effectiveness of the individual GAPrompt components. The results are generated by LlaMa2-70B with a temperature value of 0.01 before the ensembling stage.

CoT [23] and AutoCoT [50], demonstrating that the few-shot step-by-step demonstrations are the most important factor in improving the LLM inference performance.

Finally, we apply ensembling to further boost GAPrompt performance (+2.91% F1 score), minimizing the influence of randomness and hallucination of LLM generation. As detailed in Section 5.5, we set up five different temperatures for the foundation LLM generations and conduct five independent LLM inferences on the test dataset. Finally, the results of these inferences are integrated, producing the final outcome of a quantitative clinical assessment for stroke.

Table A.3 in Appendices summarizes the statistical significance test results for each component of the GAPrompt paradigm using paired sample t-tests. The results presented in this table are evaluated using a pairwise comparison based on the design of our ablation study. Starting with the fundamental LLM as the baseline, each GAPrompt component is incrementally added, and the significance of its impact is assessed. All the p_values are below 0.05, confirming that the incremental improvements achieved by adding each component to the base LLM are statistically significant. These results collectively validate the effectiveness of each component in improving the overall performance of the GAPrompt paradigm.

6. Discussion

6.1. Principal findings

In this study, we have proposed a comprehensive prompting paradigm, GAPrompt, to enhance the performance of generic foundation LLMs in completing clinical domain-specific tasks, specifically the automated clinical assessment based on EHRs. Our contributions include the following:

(1) We developed a GAPrompt paradigm comprising prompt-driven selection of LLMs, generation-augmented knowledge base construction, summary-based generation-augmented retrieval, hierarchical chain-of-thought, and ensembling. It effectively addresses the limitations of generic LLMs in clinical applications in a progressive manner.

(2) We designed a prompt-driven LLM selection process to select the foundation LLM effectively and efficiently. This is a fast and systematic

method to evaluate LLMs' applicability in fulfilling our specific task requirement through specifically designed prompting templates.

(3) Through LLM augmentation, we automatically constructed an external knowledge base, consisting of both the task-specific knowledge, *i.e.*, the NIHSS assessment criteria in this study, and the demonstrations corresponding to the EHRs. It provides clear definitions of clinical assessment criteria and enhances the performance of LLM inferencing from highly relevant, logical demonstrations.

(3) We proposed a summary-based generation-augmented retrieval (SGAR) method to retrieve task-specific knowledge and demonstrations dynamically. Through LLM-extracted summaries, this method helps the retriever focus on the critical terms of the queries, including anatomy, symptom, and inspection, to accurately retrieve the relevant assessment criteria and demonstrations.

(4) We developed a hierarchical chain-of-thought (HCoT) prompting strategy integrating macro sequential chains with micro chain-of-thought. The HCoT breaks down the complex tasks with large EHRs into progressive steps with short sentences and provides logical demonstrations for LLMs, significantly improving LLM inferencing accuracy.

(5) An ensemble strategy is utilized in the final stage to enhance the robustness and performance of GAPrompt through integrating multiple LLM generations, which helps to mitigate the randomness and hallucinations of LLMs.

The methods and models developed in this research are highly versatile and can be seamlessly applied to other clinical disease assessment tasks with comparable structures. By utilizing advanced LLMs and prompting techniques, our approach establishes a scalable framework for a wide range of applications. Future research could explore extending this methodology to multiple levels of clinical practice and automated evaluation of other diseases.

6.2. Limitations

This study highlights the effectiveness of our proposed GAPrompt in empowering the capabilities of generic foundation LLMs for clinical assessment tasks. However, it is essential to recognize certain limitations.

Firstly, our model relies on a robust foundation LLM to achieve a high accuracy, which is also the fundamental reason limiting our method's ultimate performance. As publicly available LLMs continue to evolve, we anticipate stronger LLMs and enhanced performance for our solutions.

Secondly, the current limitation on LLM context length imposes restrictions on the amount of knowledge and the number of demonstrations that can be integrated during LLM inferencing, which essentially limits LLMs' performance. Recently, research efforts have focused on extending the context length of LLMs, with expectations that this limitation will be addressed in the near future.

Lastly, the performance of LLMs is typically proportional to their size, *i.e.*, the number of parameters they contain. While lightweight models offer limited capabilities, more powerful LLMs often require substantial hardware resources, posing challenges for their practical application in real-world scenarios. Balancing performance with hardware requirements remains a crucial challenge in LLM development and deployment.

7. Conclusions

In conclusion, our study underscores the transformative potential of leveraging foundation LLMs for automating the intricate analysis of EHRs in the medical domain. Focused on quantitative clinical stroke assessment as a use case, the proposed GAPrompt, incorporating LlaMa2-70B and innovative methods including generation-augmented retrieval and hierarchical chain-of-thought, demonstrates the capacity to automatically assess and quantify EHRs. This approach not only overcomes the challenges of labor-intensive and manually conducted quantitative assessments but also extends its applicability beyond the medical domain.

CRediT authorship contribution statement

Zhanzhong Gu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Wenjing Jia:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Conceptualization. **Massimo Piccardi:** Writing – review & editing, Validation, Supervision, Methodology. **Ping Yu:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Ethics

The data we use in this study is from the previous study [21], which is approved by both the University of Technology Sydney (UTS) Human Research Ethics Committee (HREC) (EC00146) under ETH23-8230, and The Third Affiliated Hospital of Sun Yat-sen University Medical Ethics Committee under A2019-007-01. The data in this study is de-identified and secondarily used, allowing for waiver of informed consent as detailed in A2019-007-01. This study is undertaken strictly in compliance with the Australia National Statement on Ethical Conduct in Human Research (Chapter 2.1) and UTS Research Policy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The data used in this study was collected as part of the international research project RES17-1120 at the University of Technology Sydney (UTS), in collaboration with the Third Affiliated Hospital of Sun Yat-sen University, the First Affiliated Hospital of Jinan University, and the First Affiliated Hospital of Fujian Medical University. We gratefully acknowledge their contributions to the data collection and their institutional support for the original project, which enabled this research. We would like to thank all the participants of this project in UTS Global Big Data Technologies Centre and UoW Centre for Digital Transformation.

Appendix A. Distribution of the generation-augmented knowledge base

Table A.1 provides the distribution of the generation-augmented knowledge base, encompassing both task-specific knowledge (assessment criteria) and LLM-generated demonstrations (reasoning examples) related to NIHSS components. The assessment criteria are evenly distributed, with each NIHSS component represented by one entry, accounting for 6.67% of the total knowledge base. In contrast, the reasoning examples show a varied distribution, with a total of 3314 examples. Components 1b (Orientation Questions), 5a (Left Arm Motor), and 5b (Right Arm Motor) exhibit higher percentages of reasoning examples, at 7.18%, 12.91%, and 12.97%, respectively. This table highlights the comprehensive coverage of the knowledge base and the variability in demonstration examples across different NIHSS components.

Appendix B. Distribution of the assessment dataset

We show the distribution of the NIHSS quantitative assessment dataset at both micro and macro levels in Table A.2. It highlights the count and percentage of each NIHSS component across the two levels, with a total of 305 entries at the micro level and 234 at the macro level. Prominent components include 1a (Level of Consciousness), 7 (Limb Ataxia), and 9 (Best Language), which show higher percentages. This distribution provides a clear overview of how assessment criteria are represented and emphasizes the alignment between the two levels.

Table A.1
The distribution of the generation-augmented knowledge base, including the task-specific knowledge and the LLM-generated demonstrations. The count of samples related to each NIHSS component and their percentages (%) are reported.

NIHSS component	Knowledge (Assessment criteria)		Demonstrations (Reasoning examples)	
	Count	Percentage	Count	Percentage
1a	1	6.67	95	2.87
1b	1	6.67	238	7.18
1c	1	6.67	13	0.39
2	1	6.67	32	0.97
3	1	6.67	30	0.91
4	1	6.67	169	5.10
5a	1	6.67	428	12.91
5b	1	6.67	430	12.97
6a	1	6.67	361	10.89
6b	1	6.67	385	11.62
7	1	6.67	207	6.25
8	1	6.67	226	6.82
9	1	6.67	641	19.34
10	1	6.67	54	1.63
11	1	6.67	5	0.15
Total	15	100	3314	100

Table A.2
The distribution of the quantitative assessment dataset. Both micro and macro level ground truth of each NIHSS component and their corresponding percentage (%) are reported.

NIHSS component	Micro Level		Macro Level	
	Count	Percentage	Count	Percentage
1a	35	11.47	30	12.82
1b	29	9.51	20	8.55
1c	3	0.98	3	1.28
2	7	2.30	7	2.99
3	1	0.33	1	0.43
4	31	10.16	25	10.68
5a	10	3.28	10	4.27
5b	29	9.51	27	11.54
6a	11	3.61	10	4.27
6b	28	9.18	27	11.54
7	47	15.41	24	10.26
8	41	13.44	23	9.83
9	24	7.87	19	8.12
10	8	2.62	7	2.99
11	1	0.33	1	0.43
Total	305	100	234	100

Table A.3
The statistical significance test results of GAPrompt. In the table, "LLM" represents the results generated by the base LLM without any prompting strategies; "KB" refers to the Knowledge Base; "SGAR" denotes the Summary-based Generation-Augmented Retrieval; "HCoT" stands for Hierarchical Chain-of-Thought; and "ES" represents the Ensembling strategy.

GAPrompt components	LLM + KB *vs* LLM	LLM + KB +SGAR *vs* LLM + KB	LLM + KB +SGAR + HCoT *vs* LLM + KB + SGAR	LLM + KB +SGAR + HCoT + ES *vs* LLM + KB + SGAR + HCoT
*t*_value	−81.18	−56.97	−5.10	−3.02
*p*_value	3.01E−21	6.01E−19	1.31E−4	8.66E−3

Appendix C. Statistical significance test results

The table presents the statistical significance test results for each component of our GAPrompt paradigm. The *p*-values from paired sample *t*-tests are reported.

Data and code availability

The data in this study can be available upon reasonable request from the corresponding author of the existing work [21]. The code of this study can be available upon reasonable request from the corresponding author.

References

[1] Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: A machine learning approach. BMC Med Inform Decis Mak 2020;20(1):1–8.
[2] Hong C, Rush E, Liu M, Zhou D, Sun J, Sonabend A, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. NPJ Digit Med 2021;4(1):151.
[3] Xu D, Wang C, Khan A, Shang N, He Z, Gordon A, et al. Quantitative disease risk scores from EHR with applications to clinical risk stratification and genetic studies. NPJ Digit Med 2021;4(1):116.
[4] Osborne TF, Veigulis ZP, Arreola DM, Röösli E, Curtin CM. Automated EHR score to predict COVID-19 outcomes at US department of veterans affairs. PLoS One 2020;15(7):e0236554.
[5] Park E, Lee K, Han T, Nam HS, et al. Automatic grading of stroke symptoms for rapid assessment using optimized machine learning and 4-limb kinematics: Clinical validation study. J Med Internet Res 2020;22(9):e20641.
[6] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. 2018, OpenAI.
[7] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9.
[8] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.
[9] Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: A practitioner's perspective. 2022, URL https://arxiv.org/abs/2202.01875.
[10] Schaekermann M, Cai CJ, Huang AE, Sayres R. Expert discussions improve comprehension of difficult cases in medical image assessment. In: Proceedings of the 2020 CHI conference on human factors in computing systems. 2020, p. 1–13.
[11] Wang B, Min S, Deng X, Shen J, Wu Y, Zettlemoyer L, et al. Towards understanding chain-of-thought prompting: An empirical study of what matters. 2022, arXiv preprint arXiv:2212.10001.
[12] Kraljevic Z, Shek A, Bean D, Bendayan R, Teo J, Dobson R. MedGPT: Medical concept prediction from clinical narratives. 2021, arXiv preprint arXiv:2107.03134.
[13] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. 2022, arXiv preprint arXiv:2212.13138.
[14] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. 2023, arXiv preprint arXiv:2305.09617.
[15] Yang C, Wang X, Lu Y, Liu H, Le QV, Zhou D, et al. Large language models as optimizers. 2023, arXiv preprint arXiv:2309.03409.
[16] Fahes M, Vu T-H, Bursuc A, Pérez P, de Charette R. PODA: Prompt-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 18623–33.
[17] Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: Methodologies and applications. 2023, arXiv preprint arXiv:2304.14670.
[18] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 2020;33:9459–74.
[19] Mao Y, He P, Liu X, Shen Y, Gao J, Han J, et al. Generation-augmented retrieval for open-domain question answering. 2020, arXiv preprint arXiv:2009.08553.
[20] Brott T, Marler JR, Olinger CP, Adams Jr HP, Tomsick T, Barsan WG, et al. Measurements of acute cerebral infarction: Lesion size by computed tomography. Stroke 1989;20(7):871–5.
[21] Gu Z, He X, Yu P, Jia W, Yang X, Peng G, et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model. Artif Intell Med 2024;102822.
[22] Langchain-ai. Langchain. 2022, https://github.com/langchain-ai/langchain.
[23] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824–37.

[24] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations. Online: Association for Computational Linguistics; 2020, p. 38–45, URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[25] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021, arXiv preprint arXiv:2108.07258.
[26] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. 2023, arXiv preprint arXiv:2303.08774.
[27] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. 2023, arXiv preprint arXiv:2307.09288.
[28] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit Heal 2023;2(2):e0000198.
[29] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. 2023, arXiv preprint arXiv:2303.13375.
[30] Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. 2023, arXiv preprint arXiv:2311.16452.
[31] Liu Z, Li Y, Shu P, Zhong A, Yang L, Ju C, et al. Radiology-llama2: Best-in-class large language model for radiology. 2023, arXiv preprint arXiv:2309.06419.
[32] Liu S, Fang W, Lu Y, Zhang Q, Zhang H, Xie Z. RTLCoder: Outperforming GPT-3.5 in design RTL generation with our open-source dataset and lightweight solution. 2023, arXiv preprint arXiv:2312.08617.
[33] Workshop B, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, et al. Bloom: A 176b-parameter open-access multilingual language model. 2022, arXiv preprint arXiv:2211.05100.
[34] Institutes TI. Falcon. 2023, https://huggingface.co/tiiuae/falcon-40b-instruct.
[35] Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Alpaca: A strong, replicable instruction-following model. Stanf Cent Res Found Model 2023;3(6):7, https://crfm.stanford.edu/2023/03/13/alpaca.html.
[36] Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. 2023, arXiv preprint arXiv:2304.08247.
[37] Baichuan. Baichuan 2: Open large-scale language models. 2023, arXiv preprint arXiv:2309.10305, URL https://arxiv.org/abs/2309.10305.
[38] Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. 2023, arXiv preprint arXiv:2309.16609.
[39] Inc. XT. XVERSE. 2023, https://github.com/xverse-ai/XVERSE-65B.
[40] Xiong H, Wang S, Zhu Y, Zhao Z, Liu Y, Wang Q, et al. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. 2023, arXiv preprint arXiv:2304.01097.
[41] Wang H, Liu C, Xi N, Qiang Z, Zhao S, Qin B, et al. HuaTuo: Tuning llama model with Chinese medical knowledge. 2023, arXiv:2304.06975.
[42] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. 2020, arXiv preprint arXiv:2009.03300.
[43] Austin J, Odena A, Nye M, Bosma M, Michalewski H, Dohan D, et al. Program synthesis with large language models. 2021, arXiv preprint arXiv:2108.07732.
[44] Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, et al. Training verifiers to solve math word problems. 2021, arXiv preprint arXiv:2110.14168.
[45] Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, et al. Measuring mathematical problem solving with the math dataset. 2021, arXiv preprint arXiv:2103.03874.
[46] Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421.
[47] Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference, and learning. PMLR; 2022, p. 248–60.
[48] Shi Y, Xu S, Liu Z, Liu T, Li X, Liu N. Mededit: Model editing for medical question answering with external knowledge bases. 2023, arXiv preprint arXiv:2309.16035.
[49] Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. 2023, arXiv preprint arXiv:2302.12813.
[50] Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. 2022, arXiv preprint arXiv:2210.03493.
[51] Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, et al. Self-instruct: Aligning language model with self generated instructions. 2022, arXiv preprint arXiv:2212.10560.
[52] Chen D, Fisch A, Weston J, Bordes A. Reading Wikipedia to answer open-domain questions. In: Barzilay R, Kan M-Y, editors. Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). Vancouver, Canada: Association for Computational Linguistics; 2017, p. 1870–9. http://dx.doi.org/10.18653/v1/P17-1171, URL https://aclanthology.org/P17-1171.
[53] Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. 2020, arXiv preprint arXiv:2004.04906.
[54] Abburi H, Suesserman M, Pudota N, Veeramani B, Bowen E, Bhattacharya S. Generative ai text classification using ensemble llm approaches. 2023, arXiv preprint arXiv:2309.07755.
[55] Sivarajkumar S, Wang Y. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In: AMIA annual symposium proceedings, Vol. 2022. American Medical Informatics Association; 2022, p. 972.
[56] Sentence-transformers. All-mpnet-base-v2. 2021, https://huggingface.co/sentence-transformers/all-mpnet-base-v2.
[57] Rijsbergen Cv. Information retrieval. Butterworth-Heinemann; 1979.