Enhancement of Large Language Models' Performance in Diabetes Education: Retrieval-Augmented Generation Approach

Dingqiao Wang, Jiangbo Liang, Jinguo Ye, Jingni Li, Jingpeng Li, Qikai Zhang, Qiuling Hu, Caineng Pan, Dongliang Wang, Zhong Liu, Wen Shi, Danli Shi, Fei Li, Bo Qu, Yingfeng Zheng

Submitted to: Journal of Medical Internet Research
on: March 04, 2024

Table of Contents

Enhancement of Large Language Models' Performance in Diabetes Education: Retrieval-Augmented Generation Approach

Dingqiao Wang[1*]; Jiangbo Liang[1*]; Jinguo Ye[1*]; Jingni Li[1]; Jingpeng Li[1]; Qikai Zhang[1]; Qiuling Hu[1]; Caineng Pan[1]; Dongliang Wang[1]; Zhong Liu[1]; Wen Shi[1]; Danli Shi[2]; Fei Li[1]; Bo Qu[1]; Yingfeng Zheng[1]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University GuangZhou CN
[2]Research Centre for SHARP Vision, The Hong Kong Polytechnic University Hong Kong CN
[*]these authors contributed equally

Corresponding Author:
Yingfeng Zheng
State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University
GuangZhou
CN

Abstract

Background: Large language models (LLMs) demonstrated advanced performance in processing clinical information. However, commercially available LLMs lack specialized medical knowledge and remain susceptible to generating inaccurate information. Given the need for self-management in diabetes, patients commonly seek information online. We introduce the RISE framework and evaluate its performance in enhancing LLMs to provide accurate responses to diabetes-related inquiries.

Objective: This study aimed to evaluate the potential of RISE framework, an information retrieval and augmentation tool, to improve the LLM's performance to accurately respond to diabetes-related inquiries.

Methods: The RISE, an innovative Retrieval Augmentation framework, comprises four steps: Rewriting Query, Information Retrieval, Summarization, and Execution. Using a set of 43 common diabetes-related questions, we evaluated three base LLMs (GPT-4, Anthropic Claude 2, Google Bard) and their RISE-enhanced versions. Assessments were conducted by clinicians for accuracy and comprehensiveness, and by patient for understandability.

Results: The integration of RISE significantly improved the accuracy and comprehensiveness of responses from all three based LLMs. On average, the percentage of accurate responses increased by 10.9% with RISE. The rates of accurate responses increased by 7.0% for GPT-4, 16.3% for Claude 2, and 9.3% for Google Bard. The framework also enhanced response comprehensiveness, with mean scores improving by 0.44. Understandability was also enhanced by 0.19 in average.

Conclusions: RISE significantly improves LLMs' performance in diabetes-related inquiries, enhancing accuracy, comprehensiveness, and understandability. These improvements have crucial implications for RISE's future role in patient education and chronic illness self-management, which contributes to relieving medical resource pressures and raising public awareness of medical knowledge.

(JMIR Preprints 04/03/2024:58041)
DOI: https://doi.org/10.2196/preprints.58041

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Enhancement of Large Language Models' Performance in Diabetes Education: Retrieval-Augmented Generation Approach

Competing interests:
The authors declare no competing interests.

Funding:
The National Key Research and Development Program of China (2022YFC2502800); The National Natural Science Foundation of China (NSFC grant 82171034, 81721003); The High-Level Hospital Construction Project at Zhongshan Ophthalmic Center of Sun Yat-sen University (Grants 303010303058, 303020107, 303020108). The Young Talent Support Project (2022) of the Guangzhou Association for Science and Technology.

Abstract

Background

Large language models (LLMs) demonstrated advanced performance in processing clinical information. However, commercially available LLMs lack specialized medical knowledge and remain susceptible to generating inaccurate information. Given the need for self-management in diabetes, patients commonly seek information online. We introduce the RISE framework and evaluate its performance in enhancing LLMs to provide accurate responses to diabetes-related inquiries.

Objective

This study aimed to evaluate the potential of RISE framework, an information retrieval and augmentation tool, to improve the LLM's performance to accurately and safely respond to diabetes-related inquiries.

Methods

The RISE, an innovative retrieval augmentation framework, comprises four steps: Rewriting Query, Information Retrieval, Summarization, and Execution. Using a set of 43 common diabetes-related questions, we evaluated three base LLMs (GPT-4, Anthropic Claude 2, Google Bard) and their RISE-enhanced versions. Assessments were conducted by clinicians for accuracy and comprehensiveness, and by patients for understandability.

Results

The integration of RISE significantly improved the accuracy and comprehensiveness of responses from all three based LLMs. On average, the percentage of accurate responses increased by 12% (122 - 107/129) with RISE. Specifically, the rates of accurate responses increased by 7% (42 - 39/43) for GPT-4, 19% (39 - 31/43) for Claude 2, and 9% (41 - 37/43) for Google Bard. The framework also enhanced response comprehensiveness, with mean scores improving by 0.44. Understandability was also enhanced by 0.19 on average. Data collection was conducted from Sept. 30, 2023, to Feb. 05, 2024.

Conclusion

RISE significantly improves LLMs' performance in responding to diabetes-related inquiries, enhancing accuracy, comprehensiveness, and understandability. These improvements have crucial implications for RISE's future role in patient education and chronic illness self-management, which contributes to relieving medical resource pressures and raising public awareness of medical knowledge.

Keywords: Large language models (LLMs); Retrieval-augmented generation (RAG); GPT-4.0; Claude-2; Google Bard; Diabetes education

Introduction

Diabetes mellitus is a chronic long-term illness that requires continual health education and assistance to improve patient outcomes [1,2]. The shortage of diabetes counselors and the limitations of traditional education methods make it challenging to address the unique requirements of each diabetic patient [3]. Large language models (LLMs), such as ChatGPT, hold significant promise in diabetes self-management and information assessment [3-8]. However, concerns exist around the accuracy and reliability of these models, mainly stemming from the variable credibility of their training data which is sourced from a wide variety of internet text and self-supervised learning [9-11]. Furthermore, LLMs may lack domain-specific knowledge, risking the production of potentially inaccurate responses [12-15].

Recent studies have primarily assessed the capabilities of LLMs in responding to diabetes-related questions, revealing limitations in their expertise in medical specialties, which remain unresolved. For example, Meo et al. [16] indicated both ChatGPT and Google Bard scored below 60% in endocrinology and diabetes. They concluded that while these AI tools show potential in academic medical education, they require more updated information in these specific medical fields. Rachel et al. [17] also highlighted the precision of chatbots in medical queries but underlined the need for further research and model development for enhanced accuracy and validation in clinical practice. Hulman et al. [18] showed that ChatGPT-generated responses could be distinguished from expert responses by 59.5%, suggesting a gap compared to expert human performance. Therefore, addressing these gaps by augmenting LLMs with more specialized knowledge and updated information is crucial for improving their role in patient understanding and management of diabetes.

In response to these unresolved challenges, our study introduces "RISE," an independent workflow designed to enhance the performance of LLMs in the medical domain by automatically retrieving real-time external knowledge. We employed LLMs with and without RISE to answer diabetes-related inquiries from patients, assessing the improvements that RISE brings to the original LLMs in terms of accuracy, comprehensiveness and understandability. Our RISE aims to bridge the knowledge gaps identified in LLMs, providing a more robust and reliable tool for addressing patient concerns about diabetes management and understanding.

The main contributions of our work are as follows:

• We introduce RISE, an innovative framework based on the RAG algorithm that enhances LLMs with real-time, domain-specific knowledge to provide accurate and comprehensive responses to diabetes-related inquiries, improving patient self-management and outcomes.

• We reduce the risk of inaccurate or irrelevant responses from LLMs by integrating local and external real-time information retrieval, enhancing model transparency by identifying source information.

• We incorporate an additional module for accuracy and safety checks before responding, ensuring that the provided information is reliable and free from harmful content.

• We validate the RISE framework through assessments by clinicians and patients to demonstrate the feasibility of adopting RISE-enhanced LLMs in diabetes management and education.

Related works

Large Language Models

LLMs, such as GPT-3 [19], GPT-4 [20], and PaLM [21], have garnered significant attention due to their exceptional language understanding and generation capabilities [22,23]. However, when applied to domain-specific tasks, particularly in the medical field, their performance may be limited by a lack of exposure to specialized knowledge and vocabulary [24-26]. Adapting LLMs for biomedical applications poses several challenges, including insufficient domain knowledge and high computational costs. As a result, only a few LLMs have been fine-tuned for medical consultation using open-source models with 6.5 -13 billion parameters, such as ChatDoctor[27] and

MedAlpaca[28]. However, this approach of fine-tuning open-source models has its limitations. Medical domain-specific models often employ relatively smaller-scale LLMs (e.g., LLaMA [27] with 7B parameters), which may result in lower accuracy and robustness, compared to GPT-4 [29]. Moreover, fine-tuning even these smaller LLMs is computationally intensive and costly [27]. The introduction of new knowledge requires complete retraining of the model, placing additional burdens on developers. Furthermore, LLMs are generally prone to hallucination, which is a challenge that fine-tuning struggles to address [30-33].

Retrieval-Augmented Generation (RAG) in Medical Question and Answer

Recent studies have explored the application of RAG [34,35] in the medical domain to enhance the performance of LLMs in question-answering tasks. These approaches enable LLMs to achieve improved performance without needing time-intensive and costly fine-tuning while facilitating timely updates without retraining the entire model.

In specialized medical domains, LLMs have been augmented with limited medical corpora to address specific areas such as liver diseases (LiVersa)[36], diffuse large B-cell lymphoma[32], and nephrology[37]. Simultaneously, in the general medical context, frameworks such as Almanac[38] and RECTIFIER[39] have been proposed to integrate LLMs with medical guidelines and treatment recommendations.

Despite their potential, these approaches also present several limitations. The effectiveness of RAG-based models largely depends on the quality and currency of the utilized data sources. The previous studies typically rely on fixed and related smaller knowledge bases, such as Wikipedia or guidance documents, thereby limiting their effectiveness in specialized medical domains [40,41]. Outdated or incorrect information can result in inaccurate or misleading outputs. Furthermore, retrieval errors or the inclusion of biased and unsafe content inherent in the LLMs, without further filtering, may lead to inaccuracies in the generated output, potentially misleading patients.

Our RISE framework addresses these limitations by comminating with local and internet-based knowledge sources, curated explicitly from over 200 reputable academic websites, ensuring access to a wide range of up-to-date clinical evidence. Moreover, we incorporate additional fact-checking and safety check modules before responding. By prioritizing the accuracy and safety of the retrieved information, our framework offers a more reliable and secure pathway for answering clinical questions, significantly reducing the risk of misleading patients.

Methods

Framework of RISE

Our study introduced the RISE (Retrieval and Information Augmentation for Enhanced Medical Question Answering) framework, an innovative approach designed to improve the performance of medical question answering of LLMs. Our novel algorithm derives from RAG [42-44], which retrieves pertinent information from local databases or external knowledge from academic websites. Our RISE is a standalone framework comprising four steps. (**Figure 1 and Appendix 1**)

Rewriting Query

The first step in the RISE framework involves rewriting the original query using advanced LLMs, including GPT-4, Claude2, and Google Bard (Alphabet Inc.; subsequently rebranded as Gemini). This process aims to enhance the query by correcting spelling errors, expanding abbreviations, and incorporating synonyms, thereby broadening the scope of potential results.

Information Retrieval

Relevant information is retrieved from a local vectorized database and external knowledge sources. The rewritten query is embedded in the same vector space as the database, and the FAISS algorithm is used for similarity search to find the top 5 most pertinent documents (retriever=vectorstore.as_retriever (search_type="similarity", search_kwargs={"k": 5}), results =

retriever.invoke (query)). If no results are found locally, external knowledge is sourced from academic websites (over 200), ensuring that all information adheres to stringent academic and research standards.

Summarization

The third step involves condensing the retrieved information into a concise and understandable format by prompt. This step also includes fact-checking and safety checks to ensure accuracy and reduce harmful content.

Fact-checking is performed in two steps. First, before the fact-checking process, the retrieved raw text and the question are input, and the retrieved text is broken down into multiple claims. Second, these claims and the question are input, allowing the model to self-check which claims are confirmed using external knowledge sources. The model then returns the verified claims as the final summarization text.

The safety check process employs a set of 24 rules to restrict and filter the content, ensuring the generated responses are safe and appropriate. The model is prompted with the instruction: "Your answer must adhere to the following rules: {rules}.".

Execution

The final step involves presenting the summarized information and prompts to the LLMs to generate the final answer for the user. The prompt instructs: "Use the following pieces of context to answer the question at the end. Note that your response should be as brief as possible and no more than 300 words. If you don't know the answer, just say that you don't know, don't try to make up an answer.".

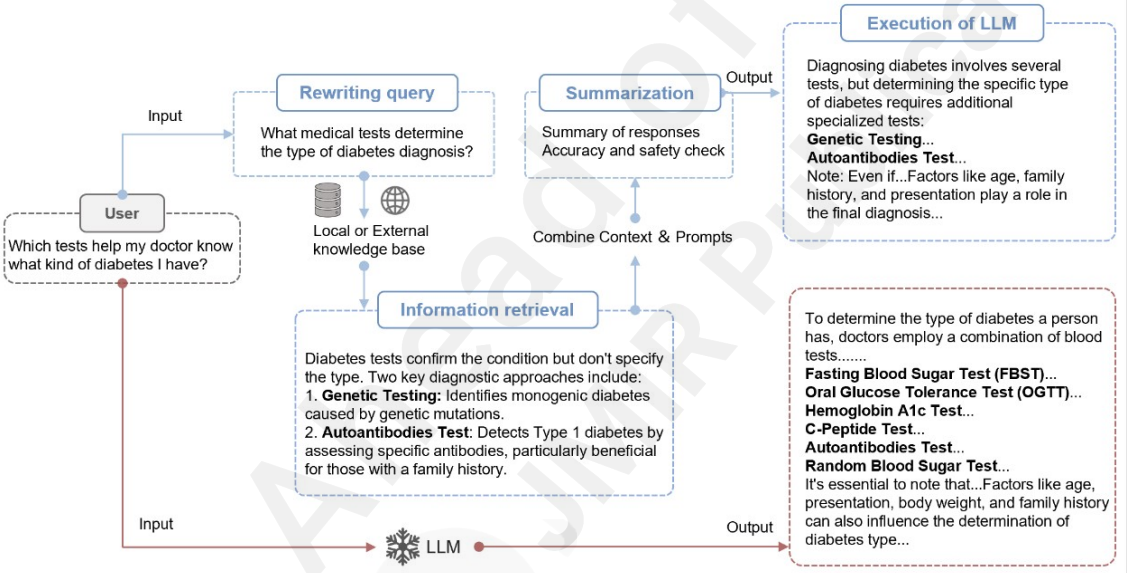


Figure 1: Comparing responses of LLMs before and after "RISE" integration. Red bars: Response from base LLMs without RISE framework. Blue bars: Overview of RISE framework and query response after integration with RISE. The framework of RISE: ① Rewriting Query: Improve query accuracy and relevance using Large Language Models (LLMs). ② Information Retrieval: Search for relevant information using the revised query from the local dataset and external knowledge base. ③ Summarization: Summarize retrieved information into concise key points, combined with fact-checking and safety checks. ④ Execution: LLMs take action based on summarized information. (See **Appendix 1** for implementation details)

Local Database

A local database of diabetes-related information was created to provide domain-specific knowledge for the RISE framework. PubMed Central (PMC) [45] was utilized to acquire a corpus of scientific papers and clinical practice guidelines relevant to diabetes. The database covers various aspects of

diabetes, including pathophysiology, diagnosis, treatment, management, and patient education, rather than answering specific questions used in the evaluation. The retrieved documents comprise over 600 full-text articles.

The retrieved documents were then preprocessed to remove potentially unstructured or noisy information, such as figures, tables, references, and author disclosures. After cleaning each document, the CharacterTextSplitter function from Langchain was used to divide the documents into smaller fragments. We then employed the OpenAI model Text-Embedding-ADA-002 as an embedding function to generate embeddings for each fragment in FAISS using the function "db=FAISS.from_documents (docs, embeddings)", where "docs" refers to the document fragments and "embeddings" refers to the Text-Embedding-ADA-002 model. The resulting index was saved locally for continuous access and retrieval using the function "db.save_local ("faiss_index")".

When a user submitted a question, the rewriting query was transformed into an embedding vector and compared to the database embeddings using cosine similarity. The top k=5 document segments with the highest similarity scores were retrieved and used as the knowledge context for the user's query. A sample of the dataset and related code are provided in **Appendices 1 and 2**

Study design

The current study was conducted at Sun Yat-sen University from Sept 25, 2023, to Feb. 30, 2024. The 43 diabetes-related questions were selected from the National Institute of Diabetes and Digestive and Kidney Diseases website [46] across the following five domains: Concepts of Diabetes, Symptoms & Causes, Diabetes Tests & Diagnosis, Managing Diabetes, and Prevention. The questions aimed to cover topics commonly asked by the public and patients regarding diabetes care.

Respond generation: We prepared a set of 43 diabetes-related inquiries to be posed to 3 base language models - GPT-4, Anthropic Claude 2, and Google Bard - as well as their respective versions enhanced by RISE. In total, there were 6 models involved, with an enhanced version corresponding to each base model. From Sept 30, 2023, to Feb. 05, 2024, we independently fed the entire set of 43 inquiries into each of the 6 models, treating each question as a separate input and resetting the conversation between queries to minimize bias.

Model responses were evaluated in a blinded, randomized manner through two aspects - first by clinician assessment focusing on accuracy and comprehensiveness, and then by patient evaluation of understandability. The evaluation process involved clinicians with over 5 years of experience in general medicine. The responses from all six models were shuffled randomly into six different rounds. To remove potential model indicators from responses, they were transformed into plain text before being distributed to three clinicians and three diabetes patients. They all analyzed responses across six rounds spaced 48 hours apart to eliminate confounding. **(Figure 2)** Responses for each model and raw scores for evaluation are provided in **Appendix 3**.

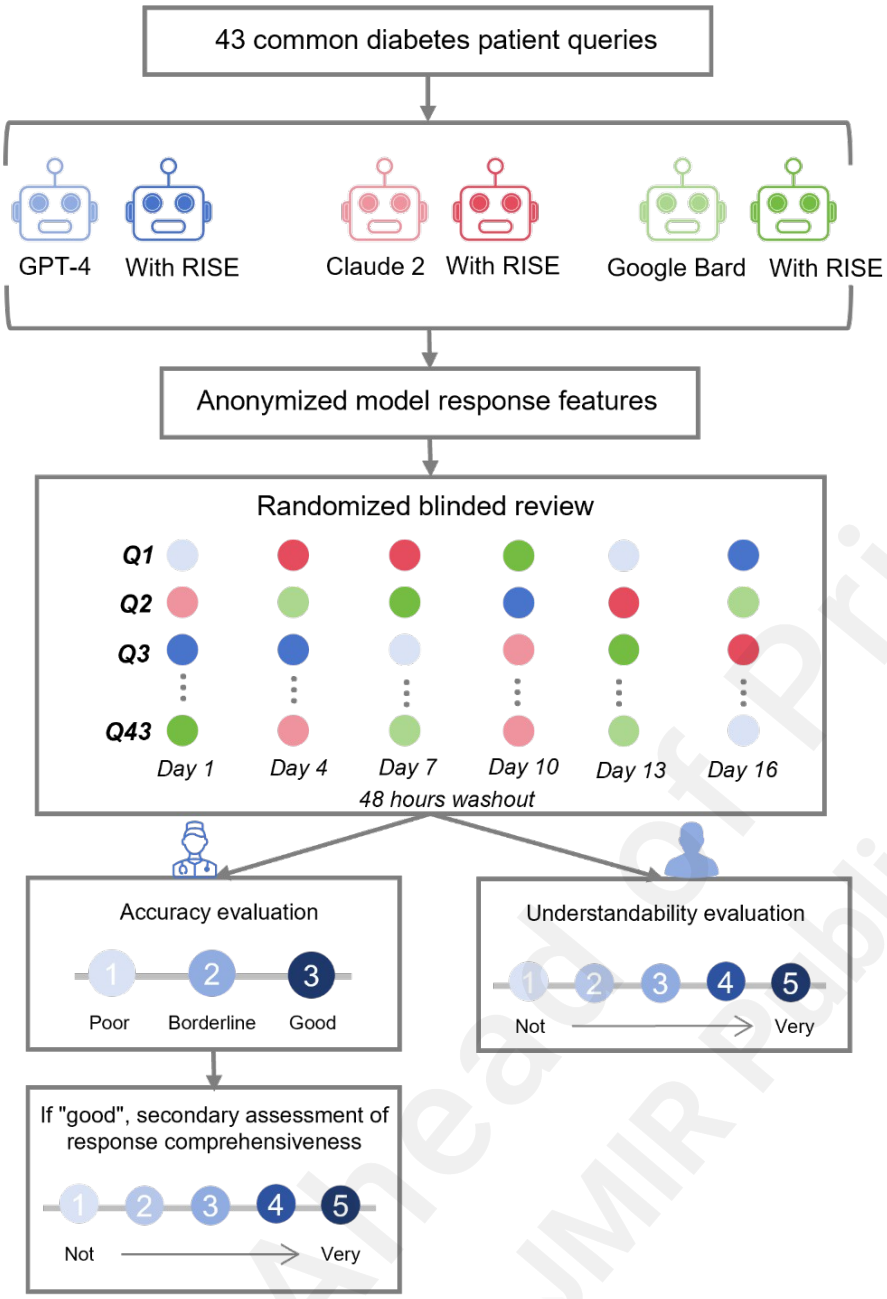


Figure 2: Flowchart of overall study design. The study evaluates the performance of three publicly available language models (LLMs) and their RISE-enhanced versions in addressing common diabetes-related inquiries. The evaluation is conducted from both the clinicians' and diabetic patients' perspectives. Clinicians evaluate the accuracy and comprehensiveness of responses. Patients assess the understandability.

Accuracy evaluation

We conducted an accuracy evaluation for each response by assigning scores and ratings. A "Poor" rating received 1 point, "Borderline" received 2 points, and "Good" received 3 points. Each response underwent assessment by three clinicians. For scoring, the score for each question is the sum of the score assigned by the three graders, with a maximum possible score of 9. For rating, we employed a majority consensus method among the three clinicians. A response was considered "Good" only if more than two clinicians rated it as such. In cases where the three clinicians provided differing ratings, we implemented a stringent strategy by giving the response the lowest mark (i.e., "poor"). The accuracy rate is defined as the proportion of responses with a final rating of "Good".
The accuracy scoring criteria include: 1) "Poor" indicating replies containing mistakes that might considerably mislead patients and potentially result in damage; 2) "Borderline" assigned to answers with potential inaccuracies but unlikely to misguide or damage patients; 3) "Good" reserved for replies without errors.

Comprehensiveness evaluation

For replies that obtained a "good" rating by majority consensus, the clinicians further evaluated the comprehensiveness of responses. A five-point scale was utilized: 1) "not comprehensive" for reactions critically missing information (1 point); 2) "slightly comprehensive" for replies with limited but primary details (2 points); 3) "moderately comprehensive" for reactions providing more than half of the essential information (3 points); 4) "comprehensive" for reactions covering most critical points (4 points); 5) "very comprehensive" or reactions giving comprehensive information (5 points). For each response, the average score was calculated by the mean of the scores assigned by the three clinicians.

Understandability evaluation

Three diabetic patients conducted an evaluation of response understandability. A five-point scale different from comprehensiveness evaluation was utilized: 1) "very Poor" for responses difficult to understand or completely irrelevant (1 point); 2) "poor" for responses somewhat difficult to understand or partially irrelevant (2 points); 3) "average" for responses generally understandable, but requiring some effort or having minor ambiguities (3 points); 4) "good" for responses most of which are easily understandable with very few unclear parts (4 points); 5) "Excellent" for responses very clear and easy to understand, fully meeting the reader's needs (5 points). For each response, the average score for understandability was calculated based on the score given by each clinician.

Statistical analysis

Statistical analyses were utilized SPSS 22.0 (SPSS Inc., USA). Normal distribution was assessed with Kolmogorov–Smirnov test. Our data were found not to follow a normal distribution, $P<0.001$. Group comparisons used the Wilcoxon signed-rank test for accuracy, comprehensiveness, and understandability scores with and without RISE. For the comparison of the proportions of "good", "borderline", and "poor" ratings across the models, the chi-square test was used. *P*-values < 0.05 were regarded as significant.

Results

Accuracy evaluation

We evaluated three LLMs and their RISE-enhanced versions for answering diabetes-related questions. As shown in **Figure 3**, the average accuracy scores of all three original models increased

substantially with the RISE enhancement. Specifically, the accuracy scores improved from 8.72 (SD 0.70) to 8.91 (SD 0.37) (P = .09) for GPT-4 after applying RISE, 8.09 (SD 1.23) to 8.65 (SD 0.65) (P = .001) for Claude, and 8.37 (SD 1.36) to 8.86 (SD 0.47) (P = .01) for Bard. (Maximum score per response is 9 points)

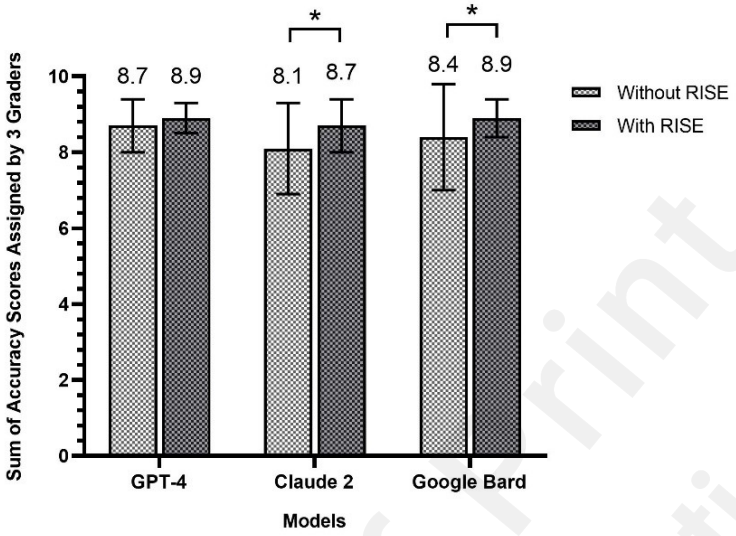


Figure 3: Average Scores of Responses from Large Language Models. Answers from each model were scored 1-3 points by 3 clinicians. The maximum score for each response is 9 points. An asterisk (*) denotes statistical significance at p < 0.05. Model Call Dates: Sept 30, 2023, to Feb. 05, 2024.

We further evaluated the percentage rated as "Good", representing accuracy, of the LLMs with and without RISE. (**Figure 4**) The results showed increased accuracy after incorporating RISE across all models. Specifically, after the incorporation of RISE, the proportion of accuracy responses for GPT-4 increased from 91% (39/43) to 98% (42/43), for Claude2 from 72% (31/43) to 91% (39/43), and for Bard from 86% (37/43) to 95% (41/43). Furthermore, GPT-4 enhanced by RISE exhibited the highest accuracy rates, reaching 98% (42/43). In addition, **Table 1** presents the accuracy of the models across 5 domains. All six models achieved the highest accuracy, reaching 100% (16/16), in the "Preventing Diabetes Problems" domain. However, in the "Concepts of Diabetes" and "Symptoms & Causes" domains, the models exhibited a relatively lower average accuracy.

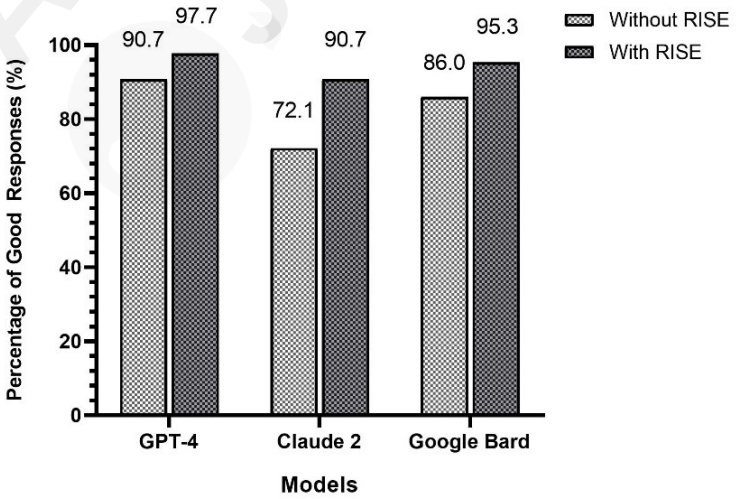


Figure 4. Accuracy rates (Proportion of "Good" Responses) of Large Language Models. Model Call Dates: Sept 30, 2023, to Feb. 05, 2024.

Table 1. Accuracy of model response across five diabetes educational domains.

n	No.	GPT-4, n (%)		Claude, n (%)		Bard, n (%)	
		Without RISE	With RISE	Without RISE	With RISE	Without RISE	With RISE
pts of Diabetes	4	2 (50.0)	4 (100)	2 (50.0)	2 (50.0)	2 (50.0)	4 (100)
oms & Causes	4	3 (75.0)	3 (75.0)	1 (25.0)	3 (75.0)	2 (50.0)	4 (100)
es Tests & Diagnosis	4	4 (100)	4 (100)	3 (75.0)	4 (100)	3 (75.0)	3 (75.0)
ing Diabetes	15	14 (93.3)	15 (100)	9 (60.0)	14 (93.3)	14 (93.3)	14 (93.3)
ting Diabetes Problems	16	16 (100)	16 (100)	16 (100)	16 (100)	16 (100)	16 (100)
	43	39 (90.7)	42 (97.7)	31 (72.1)	39 (90.7)	37 (86.0)	41 (95.3)

Accuracy indicates the percentage rated as "Good" in accuracy evaluation.

Comprehensiveness evaluation

The study also assessed the comprehensiveness of model responses through 1-5 points rating scale by 3 clinicians for the responses rated as "good". **(Table 2)** The results revealed that the integration of RISE led to a decrease in the number of responses with scores lower than 3 and an increase in the number of responses with higher scores of (4,5]. For instance, after incorporation of RISE, the number of responses scoring (1, 2] and (2, 3] reduced from 3 to 0 for GPT-4, from 6 to 3 for Claude, and from 8 to 2 for Bard. Additionally, the number of responses scoring (4, 5] increased from 19 to 38 for GPT-4, from 9 to 24 for Claude, and from 13 to 18 for Bard.

Furthermore, the average scores for comprehensiveness also improved significantly after integrating RISE. GPT-4's average score increased from 4.14 (SD 0.72) to 4.69 (SD 0.39) ($P < .001$), Claude increased from 3.79 (SD 0.78) to 4.2 (SD 0.60) ($P = .002$), and Bard increased from 3.73 (SD 0.80) to 4.10 (SD 0.62) ($P = .001$). Among the three models, GPT-4 consistently achieved the highest scores for comprehensiveness both before and after the integration of RISE, with scores of 4.14 (SD 0.72) and 4.69 (SD 0.39) respectively.

Table 2. Comprehensiveness evaluation for responses of large language models with and without RISE.

Score Range	GPT-4, n (%)		Claude, n (%)		Bard, n (%)	
	Without RISE (n = 39)	With RISE (n = 42)	Without RISE (n = 31)	With RISE (n = 39)	Without RISE (n = 37)	With RISE (n = 41)
(1, 2]	2 (5.1)	0 (0)	2 (6.5)	0 (0)	2 (5.4)	0 (0)
(2, 3]	1 (2.6)	0 (0)	4 (12.9)	3 (7.7)	6 (16.2)	2 (4.9)
(3, 4]	17 (43.6)	4 (9.5)	16 (51.6)	12 (30.8)	16 (43.2)	21 (51.2)
(4, 5]	19 (48.7)	38 (90.5)	9 (29.0)	24 (61.5)	13 (35.1)	18 (43.9)
Score (Mean ± SD)	4.14 ± 0.72	4.69 ± 0.39	3.79 ± 0.78	4.20 ± 0.60	3.73 ± 0.80	4.10 ± 0.62

For responses rated as "good" by most graders, comprehensiveness was further evaluated.

Understandability evaluation

In addition to assessing the accuracy and comprehensiveness of model responses by clinicians, this study also evaluated the public's understanding of responses. **(Table 3)** Three diabetes patients rated the understandability on a scale of 1 to 5. The results indicated the integration of RISE led to a decrease in the number of responses with scores lower than 4 and an increase in the number of responses with scores of (4, 5]. Specifically, after incorporation of RISE, the number of responses scoring lower than 4 was reduced from 15 to 4 for GPT-4, 27 to 24 for Claude, and 31 to 21 for Bard. Additionally, the number of responses with higher scores of (4, 5] increased from 18 to 39 for GPT-4, 16 to 19 for Claude, and 12 to 22 for Bard after incorporation of RISE.

Furthermore, the average scores for understandability also improved after RISE integration. GPT-4's average score increased from 4.32 (SD 0.61) to 4.64 (SD 0.51) ($P < .001$), Claude improved from 4.01 (SD 0.73) to 4.07 (SD 0.74) ($P = .31$), and Bard elevated from 3.96 (SD 0.86) to 4.16 (SD 0.82) ($P = .002$). Among the three models, GPT-4 consistently exhibited the highest understandability scores both before and after RISE integration, with scores of 4.32 (SD 0.61) and 4.64 (SD 0.51), respectively.

Table 3. Evaluation of public understandability in responses from large language models with and without RISE.

Score Range	GPT-4, n (%)		Claude, n (%)		Bard, n (%)	
	Without RISE	With RISE	Without RISE	With RISE	Without RISE	With RISE
(1, 2]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
(2, 3]	0 (0)	0 (0)	1 (2.3)	2 (4.7)	1 (2.3)	1 (2.3)
(3, 4]	15 (34.9)	4 (9.3)	26 (60.5)	22 (51.2)	30 (69.8)	20 (46.5)
(4, 5]	18 (65.1)	39 (90.7)	16 (37.2)	19 (44.2)	12 (27.9)	22 (51.2)
Score (Mean ± SD)	4.32 ± 0.61	4.64 ± 0.51	4.01 ± 0.73	4.07 ± 0.74	3.96 ± 0.86	4.16 ± 0.82

Understandability evaluation was conducted for all responses.

Discussion

Principal Findings

Considering the prevalence of T2DM as a major public health concern, particularly in light of the widespread dependence of patients on online resources for health-related information, this study introduces RISE workflow (Retrieval Augmented Generation Models) to enhance the performance of LLMs as timely and relevant diabetes education tools [43,47,48]. Our findings demonstrate that RISE significantly improves the accuracy and comprehensiveness of LLM responses to patient queries about diabetes management and care. On average, the percentage of accurate responses increased by 12% (122 - 107/129) with RISE, with rates increasing by 7% (42 - 39/43) for GPT-4, 19% (39 - 31/43) for Claude 2, and 9% (41 - 37/43) for Google Bard. The framework also enhanced response comprehensiveness and understandability, improving mean scores by 0.44 and 0.19, respectively.

Comparison to Prior Work

Previous studies have also applied LLMs in diabetes management and education. A study by Sun et al. [7] found that 74.5% of GPT-4's answers accurately responded to 200 frequently asked questions on diabetes management education. Carlos et al. [49] showed ChatGPT could correctly answer 98.5% of patient questions about type 2 diabetes, and the 1.5% inappropriate response needs to be improved. These findings were consistent with our results before integrating RISE, showing 90% (39/43) accuracy for base GPT-4 in responding to diabetes questions. Although most information provided by advanced LLMs may be correct, it's essential to realize that even small mistakes can potentially cause significant problems, especially with medical scenarios. Even minimal misinformation can lead to misconceptions, which might inadvertently delay treatment. Thus, minimizing potential errors and improving accuracy and validation are required before considering LLMs integration into patient diabetes care.

RAG has shown promise in enhancing LLM performance [50,51], however, most current RAG approaches rely on fixed, smaller, static knowledge bases. Our results showed model responses were more specific and accurate than those generated by general LLMs after incorporating specific knowledge by RISE framework, which is consistent with previous studies. Previous studies have applied RAG in other clinical specialties, such as general medicine, hepatology, and lymphoma [44,52-54]. These studies' Knowledge bases were mainly medical texts, research papers, and disease guidelines, limiting their flexibility and generalizability. In contrast, the RISE framework utilized a local medical knowledge base from NIH (National Institutes of Health) and the dynamic, real-time retrieval of external knowledge through the latest medical guidelines, academic research papers, and reputable health websites.

Future Directions

The RISE framework demonstrates the potential of RAG in enhancing the performance of LLMs for diabetes education, and there are several promising directions for future research and development. These include creating large specialized medical knowledge bases tailored for diabetes education, integrating multimodal data such as medical images and electronic health records, and developing domain-specific retrieval and ranking algorithms for evidence-based information [55,56]. Furthermore, exploring the bilingual or multilingual potential of these chatbots, such as investigating the performance of the RISE framework when questions are asked in languages like Chinese, could expand their use in real-world clinical practice outside the English-speaking world. Another promising direction is exploring personalized RAG systems that adapt to individual patients' preferences and contexts. Ensuring RAG systems' interpretability, transparency, privacy, and security

is crucial in the medical domain.

Strengths and Limitations

The current study has several strengths. Firstly, novel RAG algorithms that effectively utilize local databases and external academic knowledge markedly improve the precision and real-time performance of responses to diabetes-related inquiries. Secondly, the RISE framework incorporates rigorous factual and safety checks for the generated outputs, ensuring reliable and secure responses.
There are some limitations in the study. RISE framework was developed and evaluated exclusively within the domain of diabetes education. The generalizability of RISE to other medical domains remains uncertain. Future investigations should extend the application of the RISE framework to diverse medical specialties. Moreover, the scope of our evaluation was limited to these predetermined queries. Future research should conduct clinical trials to assess the RISE's ability to effectively address patients' inquiries and enhance the efficiency of diabetes management in real-world clinical scenarios.

Conclusions

In conclusion, RISE framework shows promise as a safer and more reliable option for generating responses to common queries from diabetes patients. RISE significantly enhances the accuracy and comprehension of original LLM responses by retrieval of external knowledge from reliable sources. This framework can potentially be a supplementary tool to improve patient understanding and disease outcomes.

Appendices

Appendix 1: Code for the RISE framework.
Appendix 2: Sample of RISE dataset (n=50); the complete dataset is available upon request.
Appendix 3: Responses for each model and raw scores for evaluation.

References

1. Buse JB, Wexler DJ, Tsapas A, Rossing P, Mingrone G, Mathieu C, et al. 2019 Update to: Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes care. 2020 Feb;43(2):487-93. PMID: 31857443. doi: 10.2337/dci19-0066.
2. Pastors JG, Warshaw H, Daly A, Franz M, Kulkarni K. The evidence for the effectiveness of medical nutrition therapy in diabetes management. Diabetes care. 2002 Mar;25(3):608-13. PMID: 11874956. doi: 10.2337/diacare.25.3.608.
3. Gao C, Xu J, Liu Y, Yang Y. Nutrition Policy and Healthy China 2030 Building. European journal of clinical nutrition. 2021 Feb;75(2):238-46. PMID: 33219269. doi: 10.1038/s41430-020-00765-6.
4. Khan I, Agarwal R. Can ChatGPT Help in the Awareness of Diabetes? Annals of biomedical engineering. 2023 Oct;51(10):2125-9. PMID: 37648882. doi: 10.1007/s10439-023-03356-1.
5. Wang DQ, Feng LY, Ye JG, Zou JG, Zheng YF. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. 2023;2(2):e43.
6. Chlorogiannis DD, Apostolos A, Chlorogiannis A, Palaiodimos L, Giannakoulas G, Pargaonkar S, et al. The Role of ChatGPT in the Advancement of Diagnosis, Management, and Prognosis of Cardiovascular and Cerebrovascular Disease. Healthcare (Basel, Switzerland). 2023 Nov 6;11(21). PMID: 37958050. doi: 10.3390/healthcare11212906.
7. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models: Preclinical Concept Validation Study. Journal of medical Internet research. 2023 Nov 9;25:e51300. PMID: 37943581. doi: 10.2196/51300.
8. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content. Journal of medical Internet research. 2024 Apr 25;26:e55847. PMID: 38663010. doi: 10.2196/55847.
9. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023 Feb;614(7947):214-6. PMID: 36747115. doi: 10.1038/d41586-023-00340-6.
10. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatology communications. 2023 Apr 1;7(4). PMID: 36972383. doi: 10.1097/hc9.0000000000000097.
11. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. Journal of medical Internet research. 2023 May 31;25:e46924. PMID: 37256685. doi: 10.2196/46924.
12. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023 Feb;614(7947):224-6. PMID: 36737653. doi: 10.1038/d41586-023-00288-7.
13. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Critical care (London, England). 2023 Mar 21;27(1):120. PMID: 36945051. doi: 10.1186/s13054-023-04393-x.
14. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023 Sep;95:104770. PMID: 37625267. doi: 10.1016/j.ebiom.2023.104770.
15. Haddad F, Saade JS. Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study. JMIR medical education. 2024 Jan 18;10:e50842. PMID: 38236632. doi: 10.2196/50842.

16. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. J Diabetes Sci Technol. 2023 Oct 5:19322968231203987. PMID: 37798960. doi: 10.1177/19322968231203987.
17. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA internal medicine. 2023 Jun 1;183(6):589-96. PMID: 37115527. doi: 10.1001/jamainternmed.2023.1838.
18. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, et al. ChatGPT-versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. PloS one. 2023;18(8):e0290773. PMID: 37651381. doi: 10.1371/journal.pone.0290773.
19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877-901.
20. OpenAI. Gpt-4 technical report. arxiv 2303.08774. 2023;2(5).
21. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research. 2023;24(240):1-113.
22. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022;35:24824-37.
23. Shi F, Suzgun M, Freitag M, Wang X, Srivats S, Vosoughi S, et al. Language models are multilingual chain-of-thought reasoners. arxiv preprint arxiv: 2210.03057. 2022.
24. Chalkidis IJapa. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. arxiv preprint arxiv: 2304.12202. 2023.
25. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev DJapa. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. arxiv preprint arxiv: 2303.18027. 2023.
26. West CGJapa. AI and the FCI: Can ChatGPT project an understanding of introductory physics? arxiv preprint arxiv: 2303.01067. 2023.
27. Yunxiang L, Zihan L, Kai Z, Ruilong D, You Z. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:230314070. 2023.
28. 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. arXiv preprint arXiv:230408247. 2023.
29. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
30. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine. 2023;183(6):589-96. PMID: 37115527. doi:10.1001/jamainternmed.2023.1838
31. Golovneva O, Chen M, Poff S, Corredor M, Zettlemoyer L, Fazel-Zarandi M, et al. Roscoe: A suite of metrics for scoring step-by-step reasoning. arXiv preprint arXiv:221207919. 2022.
32. Soong D, Sridhar S, Si H, Wagner J-S, Sá ACC, Yu CY, et al. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. arXiv preprint arXiv:230517116. 2023.
33. Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, et al. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems. 2024;36.
34. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.
35. Guu K, Lee K, Tung Z, Pasupat P, Chang M, editors. Retrieval augmented language model

pre-training. International conference on machine learning; 2020: PMLR.
36. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation. medRxiv. 2023.
37. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina. 2024;60(3):445. PMID: 38541171.doi: 10.3390/medicina60030445.
38. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac—retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1(2):AIoa2300068. PMID: 38343631. doi:10.1056/aioa2300068.
39. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. medRxiv. 2024:2024.02. 08.24302376.
40. Yu W, editor. Retrieval-augmented generation across heterogeneous knowledge. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop; 2022.
41. Feng Z, Feng X, Zhao D, Yang M, Qin B. Retrieval-Generation Synergy Augmented Large Language Models. arXiv preprint arXiv:231005149. 2023.
42. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks.Advances in Neural Information Processing Systems. 2020;33:9459-74.
43. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. Journal of medical Internet research. 2024 Apr 17;26:e56655. PMID: 38630520. doi: 10.2196/56655.
44. Shi W, Zhuang Y, Zhu Y, Iwinski H, Wattenbarger M, Wang MD, editors. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2023.
45. NCBI.PMC. https://www.ncbi.nlm.nih.gov/pmc/.[accessed Dec.5 2023].
46. Diabetes. https://www.niddk.nih.gov/health-information/diabetes.[accessed Dec.5 2023].
47. Al-Lawati JA. Diabetes Mellitus: A Local and Global Public Health Emergency! Oman medical journal. 2017 May;32(3):177-9. PMID: 28584596. doi: 10.5001/omj.2017.34.
48. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pakistan journal of medical sciences. 2023 Mar-Apr;39(2):605-7. PMID: 36950398. doi: 10.12669/pjms.39.2.7653.
49. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, Amoro Sanchez R, Muyolema Arce V, Mustafa A, et al. The Future of Patient Education: AI-Driven Guide for Type 2 Diabetes. Cureus. 2023 Nov;15(11):e48919. PMID: 38024047. doi: 10.7759/cureus.48919.
50. Chen J, Lin H, Han X, Sun L, editors. Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence; 2024.
51. Lozano A, Fleming SL, Chiang C-C, Shah N, editors. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024; 2023: World Scientific.
52. Khene Z-E, Bigot P, Mathieu R, Rouprêt M, Bensalah KJEUO. Development of a personalized chat model based on the European Association of Urology oncology guidelines: harnessing the power of generative artificial intelligence in clinical practice. European Urology Oncology. 2024;7(1):160-2.
53. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac—retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1(2):AIoa2300068. PMID:

38343631. doi:10.1056/aioa2300068.
54. Soong D, Sridhar S, Si H, Wagner J-S, Sá ACC, Yu CY, et al. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. arxiv preprint arxiv: 2305.17116. 2023.
55. Huang Y, Huang J. A Survey on Retrieval-Augmented Text Generation for Large Language Models. arXiv preprint arXiv:240410981. 2024.
56. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and efficient foundation language models. arxiv preprint arxiv: 2302.13974. 2023.

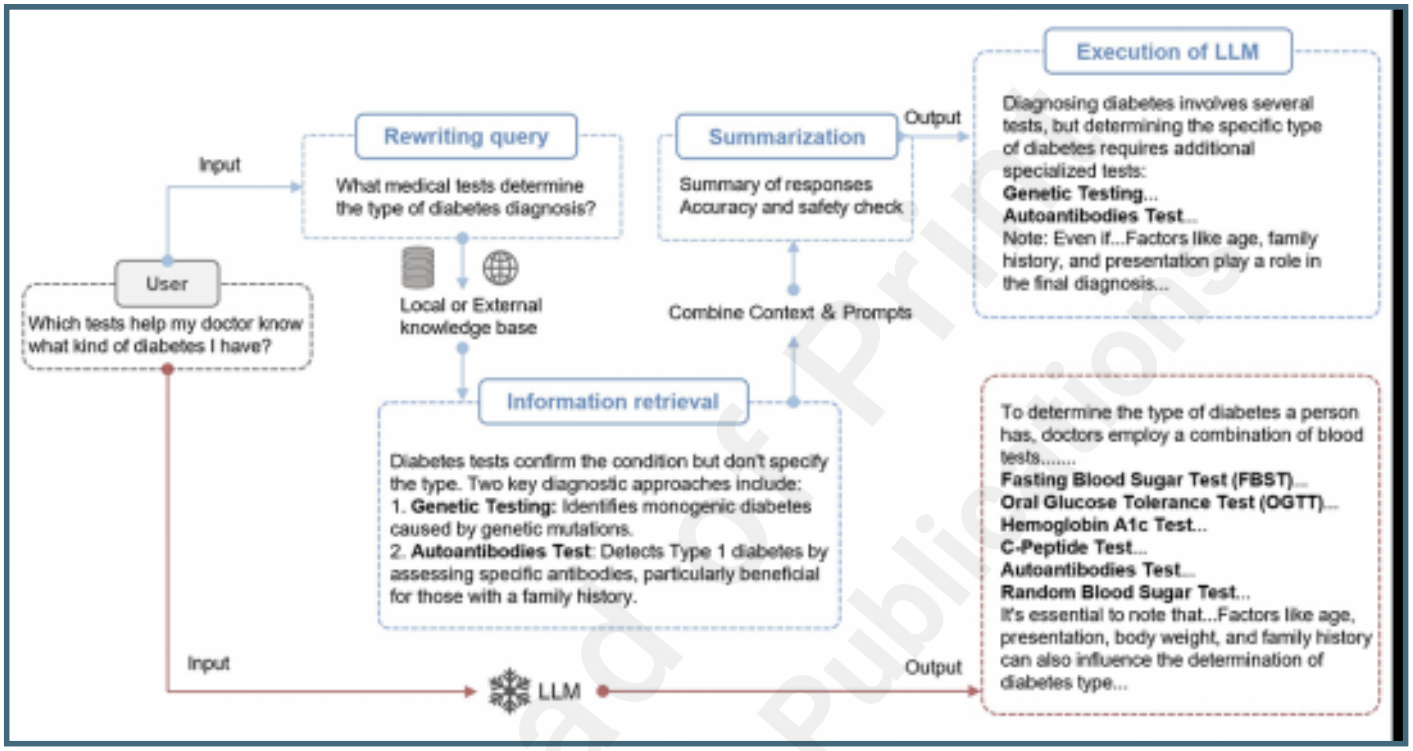
Data Availability

A subset of our dataset (n=50) is published alongside this manuscript (Appendix 2). Due to privacy and ethical considerations, the complete dataset cannot be made publicly available. However, it is available upon request. Please contact D.W. (wangdq5@mail2.sysu.edu.cn) for access to the full dataset.

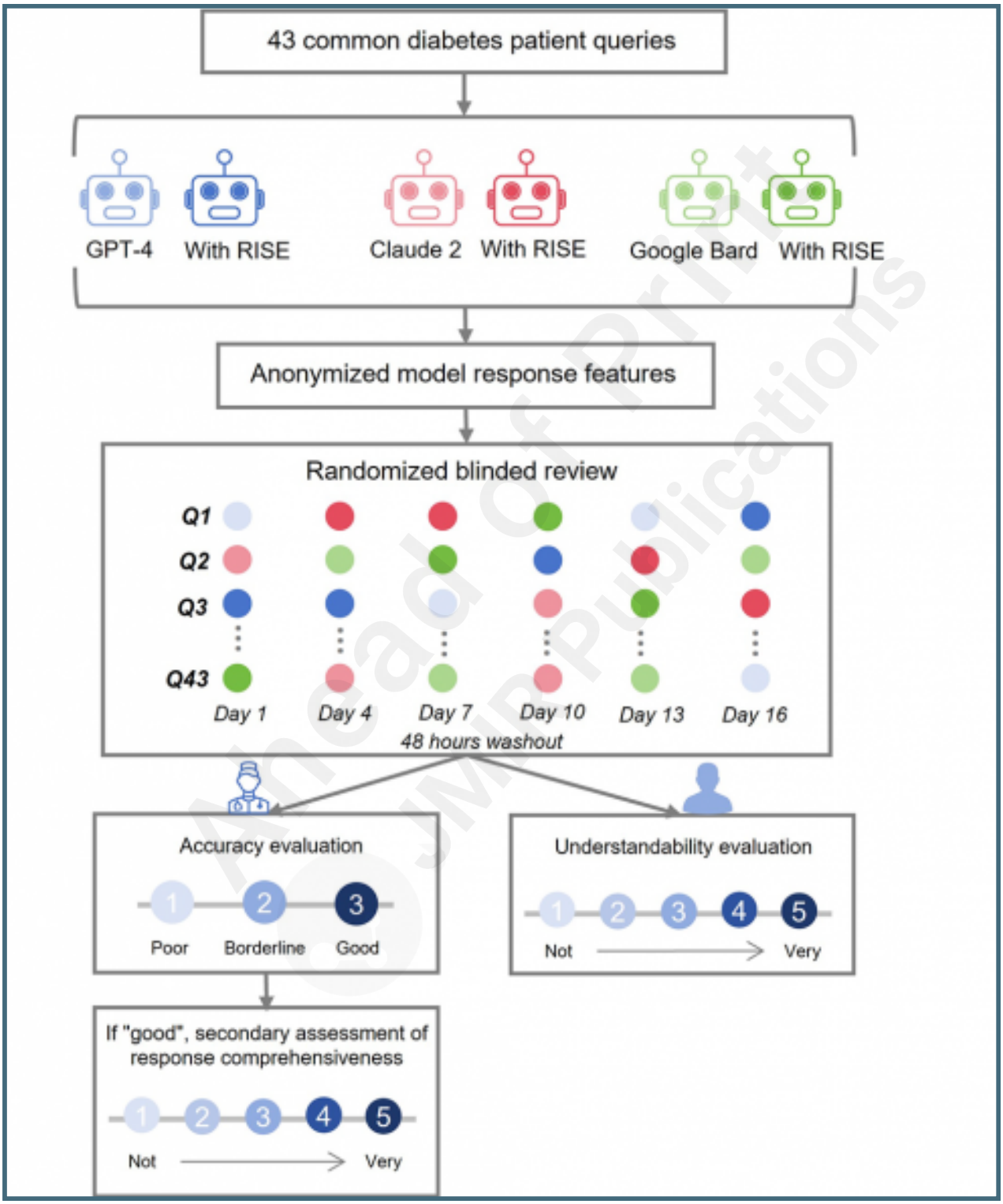
Supplementary Files

Figures

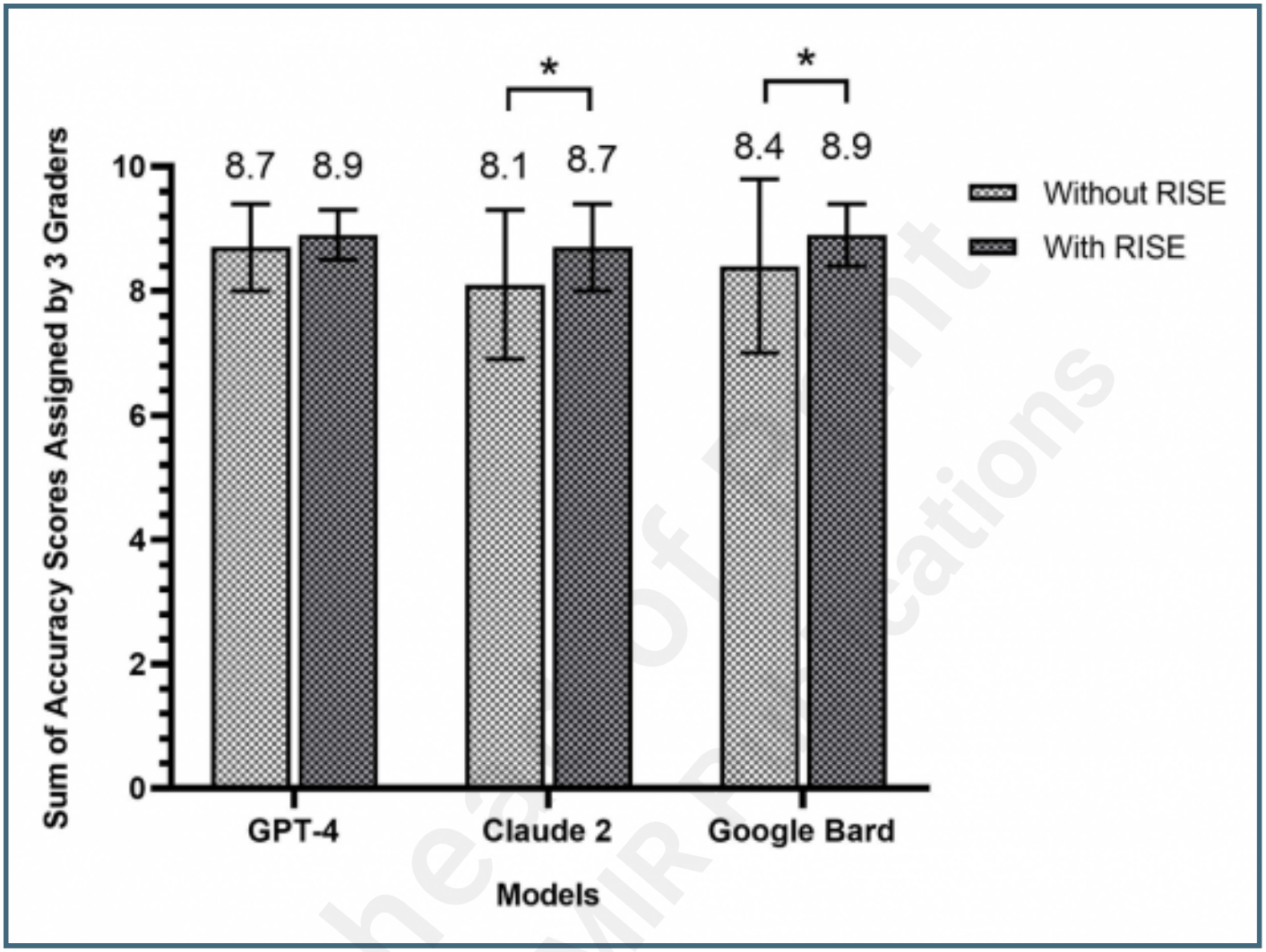
Comparing responses of LLMs before and after "RISE" integration. Red bars: Response from base LLMs without RISE framework. Blue bars: Overview of RISE framework and query response after integration with RISE. The framework of RISE: ? Rewriting Query: Improve query accuracy and relevance using Large Language Models (LLMs). ? Information Retrieval: Search for relevant information using the revised query from the local dataset and external knowledge base. ? Summarization: Summarize retrieved information into concise key points, combined with fact-checking and safety checks. ? Execution: LLMs take action based on summarized information. (See Appendix 1 for implementation details).



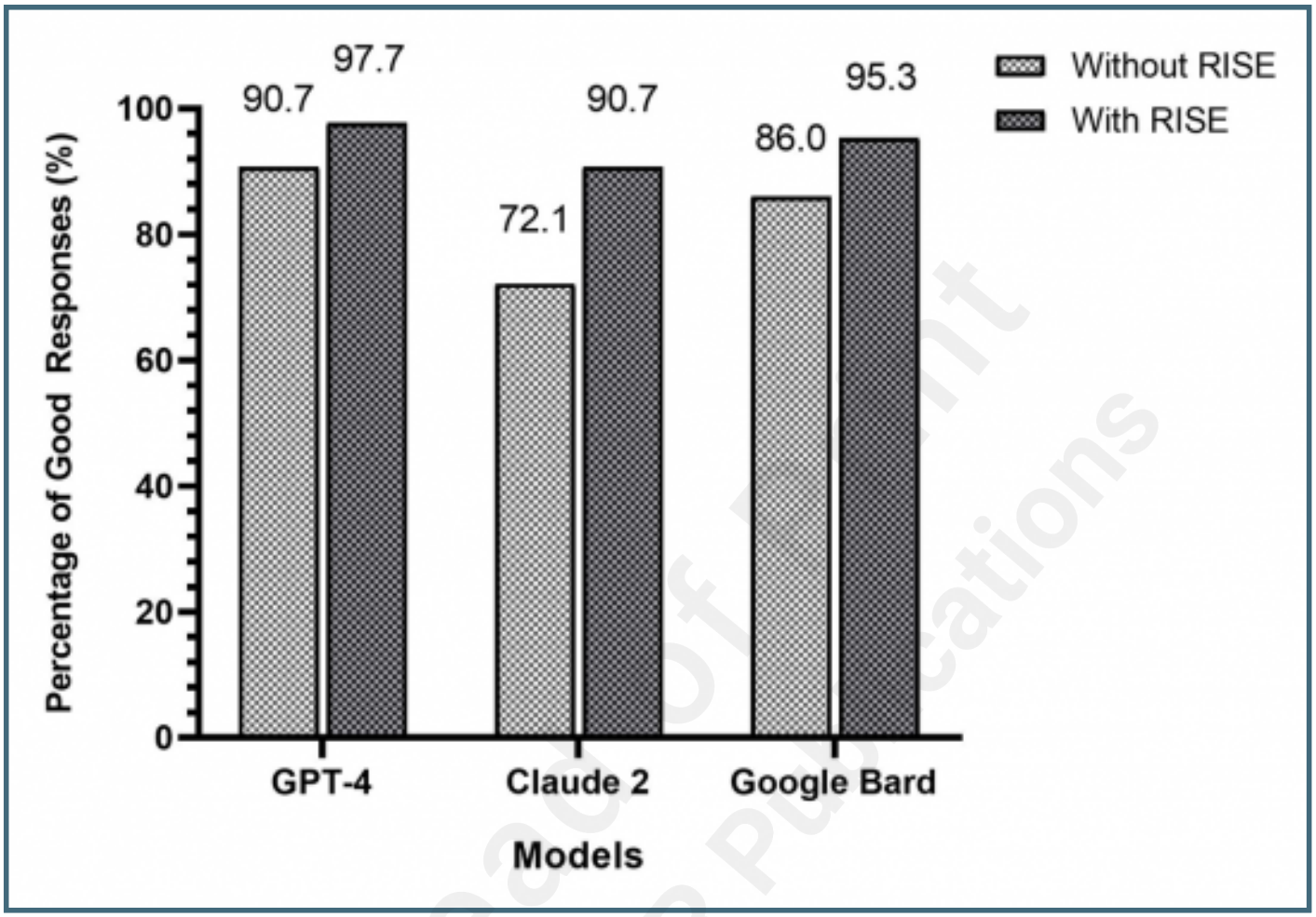
Flowchart of overall study design. The study evaluates the performance of three publicly available language models (LLMs) and their RISE-enhanced versions in addressing common diabetes-related inquiries. The evaluation is conducted from both the clinicians' and diabetic patients' perspectives. Clinicians evaluate the accuracy and comprehensiveness of responses. Patients assess the understandability.



Average Scores of Responses from Large Language Models. Answers from each model were scored 1-3 points by 3 clinicians. The maximum score for each response is 9 points. An asterisk (*) denotes statistical significance at $p < 0.05$. Model Call Dates: Sept 30, 2023, to Feb. 05, 2024.



Accuracy rates (Proportion of "Good" Responses) of Large Language Models. Model Call Dates: Sept 30, 2023, to Feb. 05, 2024.



Multimedia Appendixes

Code for the RISE framework.
URL: http://asset.jmir.pub/assets/93b5170f64ae07557c583a1630d8ad6c.zip

Sample of the RISE Diabetes Q&A Dataset (n=50).
URL: http://asset.jmir.pub/assets/b0cb336dd596292f93aecaade426d09f.xlsx

Responses for each model and raw scores for evaluation (clean).
URL: http://asset.jmir.pub/assets/0b63e54de30fd047aff60285020f78e2.pdf