OPTIMIZATIONS IN APPLIED MACHINE LEARNING
Research Article* | *Volume 5* | *Issue 5* | *Mar 2025
Received: 16 Jan 2025 | Revised: 30 Jan 2025
Accepted: 25 Feb 2025 | Published Online: 5 Mar 2025

Retrieval-augmented generation for personalized physician recommendations in online medical services: model development study

Yingbin Zheng[1], Yiwei Yan[1], Sai Chen[2], Yunping Cai[2], Kun Ren[2], Yishan Liu[3], Jiaying Zhuang[1] and Min Zhao[1*]

[1] Biomedical Big Data Center, The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China,

[2] Meteorological Disaster Prevention Technology Center, Xiamen Meteorological Bureau, Xiamen, China,

[3] School of Software Engineering, Taiyuan University of Technology, Taiyuan, China

Abstract: Web-based medical services have expanded access to healthcare through remote consultations and streamlined scheduling, but personalized physician recommendations remain limited due to reliance on manual triage. This study developed and validated a Retrieval-Augmented Generation-Based Physician Recommendation (RAGPR) model to enhance triage performance. Using 646,383 consultation records from the Internet Hospital of the First Affiliated Hospital of Xiamen University, we evaluated embedding models (FastText, SBERT, OpenAI) for clustering and classification, as well as large language models (Mistral, GPT-4o-mini, GPT-4o). Three triage staff also assessed model efficiency via questionnaires. Results showed that FastText performed poorly (F1-score 46%), while SBERT and OpenAI achieved 95% and 96%. Among LLMs, GPT-4o reached the highest F1-score (95%) with a performance rating of 4.67, followed by Mistral (94%, 4.56) and GPT-4o-mini (92%, 4.45). Considering accuracy, cost, and implementation, SBERT and Mistral were optimal. The RAGPR model offers a scalable approach to improving accuracy and personalization in online patient − physician matching.

Keywords: *large language models; mistral, SBERT; triage systems; retrieval-augmented generation-based physician recommendation; RAGPR model*

1. Introduction

Web-based medical services have significantly enhanced healthcare accessibility by improving convenience and efficiency through features such as remote consultations, streamlined scheduling, and enhanced access to medical information (1). Nevertheless, challenges persist, particularly in delivering personalized physician recommendations (2). The diverse array of medical professionals and the varying needs of patients complicate the effective identification of suitable physicians. Currently, most triage processes depend on manual recommendations made by schedulers to guide patients to the appropriate departments or practitioners (3). The increasing volume of consultations reveals the limitations of such manual methods in maintaining quality and professionalism in healthcare delivery (4). Moreover, the intermittent availability of schedulers can disrupt patient access and continuity of care, underscoring the need for a sophisticated recommendation model.

A substantial number of health-related websites now incorporate symptom checker tools (5) that offer preliminary assessments based on user inputs, employing decision trees (6) or rule-based methodologies (7, 8). Following this initial evaluation, the system proposes possible medical conditions and recommends relevant healthcare providers. Notably, machine learning algorithms, including collaborative filtering (9) and content-based filtering (10), have been investigated for their efficacy in recommending physicians by analyzing patient history, preferences, and demographic data. Advances in technology, particularly in natural language processing, present promising opportunities to utilize extensive datasets for generating tailored and precise recommendations.

The Retrieval-Augmented Generation (RAG) (11) framework presents a promising strategy for enhancing the precision and personalization of medical recommendations. Originally designed for handling fact-based inquiries within conversational models (12), RAG comprises two key components: a retriever (13) that locates relevant documents and a generator (14) that synthesizes these documents into coherent outputs. Embedding models play a crucial role in this process by providing linguistic representations that encapsulate semantic meanings in the form of numerical vectors, which are essential for retrieval systems. Simultaneously, generators, such as OpenAI's GPT series of large language models (LLMs) (15), have demonstrated significant proficiency in producing human-like text and understanding, influencing numerous natural language processing applications, including automated customer support and content generation. By integrating information retrieval with generative modeling, RAG allows systems to generate contextually rich responses that incorporate relevant external data sources (16), thus grounding responses in factual information and substantially.

2. Abbreviations:

RAG, Retrieval-Augmented Generation; LLMs, large language models; SBERT, Sentence Bidirectional Encoder Representations from Transformers; RAGPY, Retrieval-Augmented Generation-Based Physician Recommendation. improving accuracy while reducing the likelihood of misleading or erroneous outputs.

3. Methods

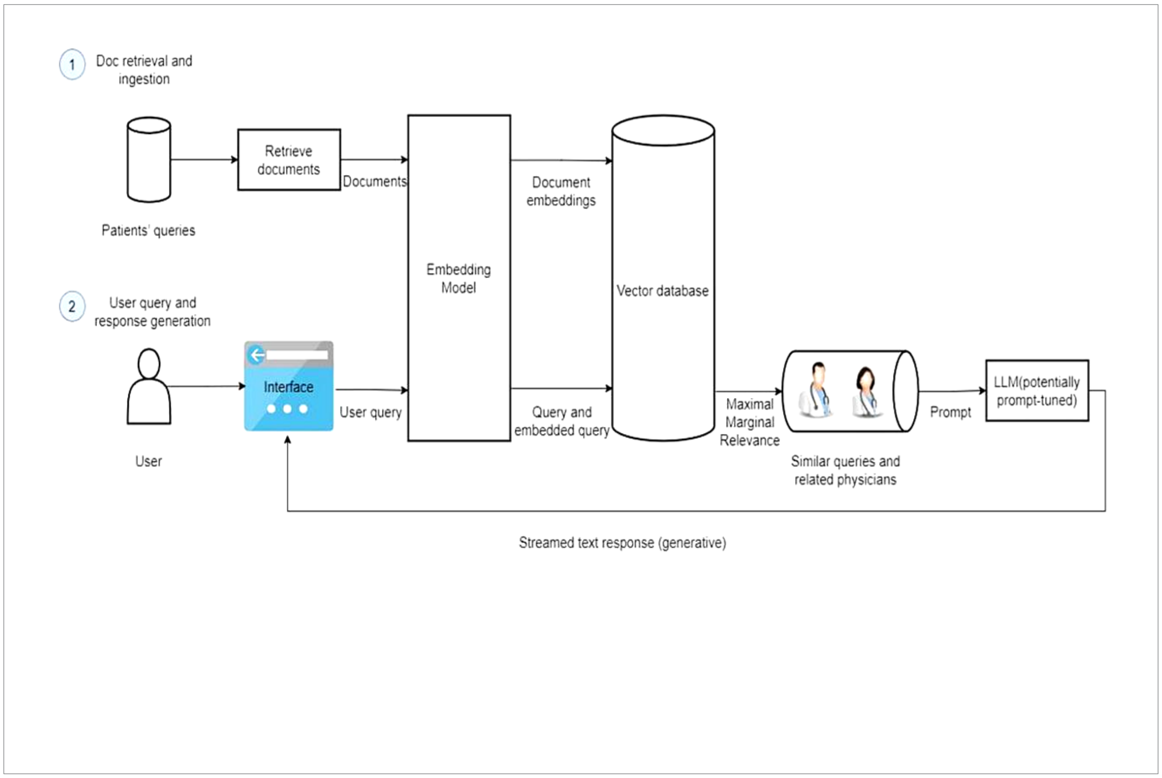
The Retrieval-Augmented Generation-Based Physician Recommendation (RAGPR) model, as illustrated in the accompanying figure, comprises two principal components: document retrieval and ingestion, and the generation of user queries and responses. A comprehensive account of the methodology is presented below (Figure 1).

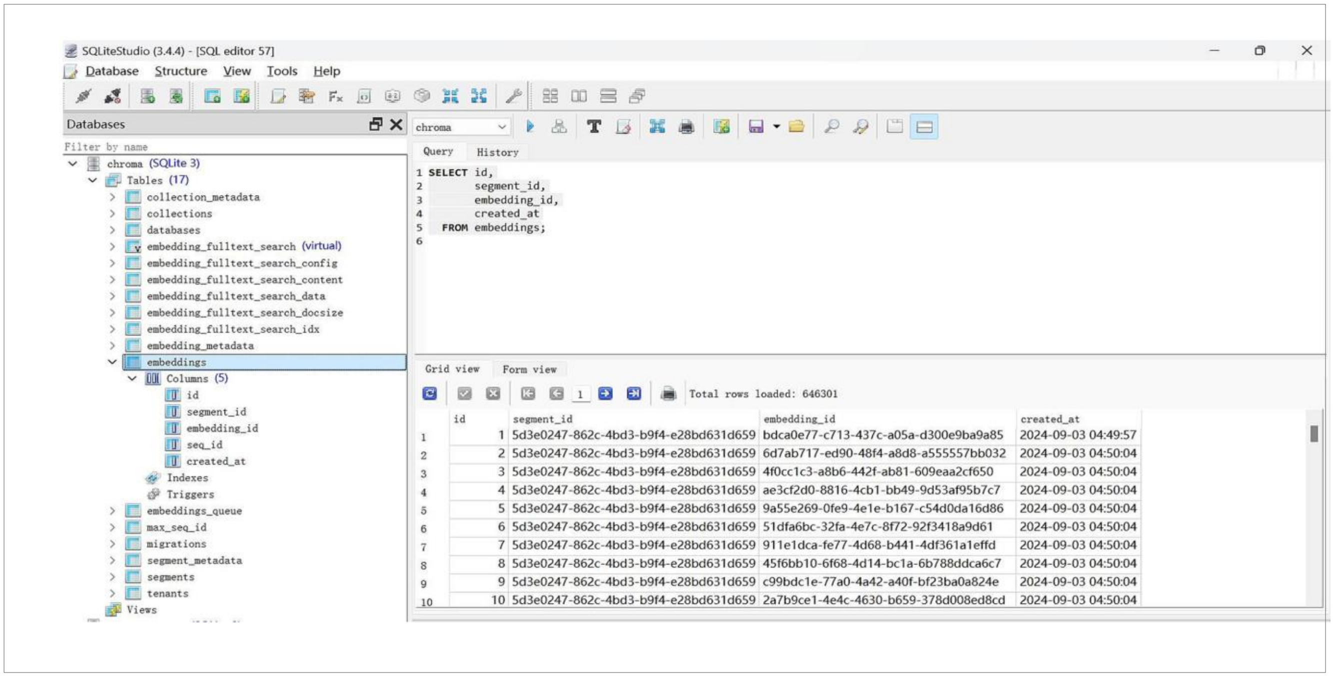
3.1 Data collection and preprocessing

The research process commenced with the collation of medical documents pertaining to patients. The training dataset was comprised of 646,383 web-based medical documents, collected from the Internet Hospital of the First Affiliated Hospital of Xiamen University. The documents spanned the years 2016 to 2023. Subsequently, data preprocessing was conducted to ensure that the data were anonymized, structured, and formatted for efficient retrieval and analysis. Each document in this dataset includes the textual query, de-identified codes for the physician and the patient, as well as information on the physician's department and response time. In contrast, the test dataset consisted of 965 web-based medical documents obtained from Hugging Face, each containing a disease label and a textual query.

3.2 Feature extraction and storage

The process of document analysis employs an embedding model that utilizes a pretrained Sentence-BERT (SBERT) model (17), specifically "distilute-base-multilingual-cased, "to transform textual information into numerical embeddings. The resulting embeddings capture the semantic essence of the text (18). The embedding process reduces the dimensionality of the data while preserving semantic similarity, thereby facilitating efficient data storage and retrieval. The resulting document embeddings are stored in Chroma (19), a specialized vector database optimized for managing high-dimensional data. This optimization allows for the rapid execution of similarity searches and serves as a repository for all vector representations of the processed documents, ensuring their accessibility for future retrieval (Figure 2).


1
Doc retrieval and ingestion
Patients' queries
Retrieve documents
Documents
Embedding Model
Document embeddings
Vector database
2
User query and response generation
User
Interface
User query
Query and embedded query
Maximal Marginal Relevance
Similar queries and related physicians
Prompt
LLM(potentially prompt-tuned)
Streamed text response (generative)



```
SELECT id,
       segment_id,
       embedding_id,
       created_at
  FROM embeddings;
```

Total rows loaded: 646301

	id	segment_id	embedding_id	created_at
1	1	5d3e0247-862c-4bd3-b9f4-e28bd631d659	bdca0e77-c713-437c-a05a-d300e9ba9a85	2024-09-03 04:49:57
2	2	5d3e0247-862c-4bd3-b9f4-e28bd631d659	6d7ab717-ed90-48f4-a8d8-a555557bb032	2024-09-03 04:50:04
3	3	5d3e0247-862c-4bd3-b9f4-e28bd631d659	4f0cc1c3-a8b6-442f-ab81-609eaa2cf650	2024-09-03 04:50:04
4	4	5d3e0247-862c-4bd3-b9f4-e28bd631d659	ae3cf2d0-8816-4cb1-bb49-9d53af95b7c7	2024-09-03 04:50:04
5	5	5d3e0247-862c-4bd3-b9f4-e28bd631d659	9a55e269-0fe9-4e1e-b167-c54d0da16d86	2024-09-03 04:50:04
6	6	5d3e0247-862c-4bd3-b9f4-e28bd631d659	51dfa6bc-32fa-4e7c-8f72-92f3418a9d61	2024-09-03 04:50:04
7	7	5d3e0247-862c-4bd3-b9f4-e28bd631d659	911e1dca-fe77-4d68-b441-4df361a1effd	2024-09-03 04:50:04
8	8	5d3e0247-862c-4bd3-b9f4-e28bd631d659	45f6bb10-6f68-4d14-bc1a-6b788ddca6c7	2024-09-03 04:50:04
9	9	5d3e0247-862c-4bd3-b9f4-e28bd631d659	c99bdc1e-77a0-4a42-a40f-bf23ba0a824e	2024-09-03 04:50:04
10	10	5d3e0247-862c-4bd3-b9f4-e28bd631d659	2a7b9ce1-4e4c-4630-b659-378d008ed8cd	2024-09-03 04:50:04

3.3 User query and response generation

The system's user interface is developed using the Vue framework (20) and connects to the backend service via a RESTful API (21). It is designed to help users clearly articulate their information needs, allowing the system to efficiently process requests. The user's input query is then transformed into an embedding using the same embedding model employed in processing the documents. This ensures compatibility and comparability between the document embeddings and the query embedding.

Maximal Marginal Relevance (MMR) is a technique employed in information retrieval to identify documents that are not only pertinent to a given query but also exhibit diversity in relation to those previously selected. The system implements MMR by comparing the embedding of the user query with the embeddings of each stored document. This method effectively reduces redundancy and enhances the coverage of various aspects of the query within the selected documents. To further improve the safety of post - retrieval recommendations, the self - reflective mechanism proposed by Zhang et al. (2024) could be referenced—this mechanism dynamically revises retrieval results through multi - dimensional validation (e.g., drug interactions, dosage limitations), providing insights for optimizing the rationality of the RAGPR model's recommendations (Zhang et al., 2024). Subsequently, the system retrieves the top 6 closely matched queries along with the corresponding physician information for the next step.

A locally constructed LLM was developed using the LLaMA(22) architecture, incorporating the Mistral-7B (23) model for pre-training parameters. This LLM is provided with a prompt that includes similar retrieved queries and related physician information. The prompt facilitates the LLM in generating coherent and contextually appropriate natural language responses. This generative capability ensures that users receive not only straightforward data retrieval but also insightful interpretations and explanations, thereby enhancing their understanding and aiding in the decision-making process. The prompt is as follows: "You are an assistant for question-answering tasks. Use the following pieces of retrieved context to recommend a department and physicians with the shorter response time. The output must be in JSON format and contain only department and physicians."

3.4 Evaluation

The evaluation of the proposed RAGPR model's effectiveness employed three key metrics: precision, recall, and F1-score. Precision is defined as the ratio of correctly identified positive samples to the total number of samples predicted as positive. This metric indicates the accuracy of the model in its positive predictions. In contrast, recall quantifies the proportion of actual positive samples that the model accurately identifies, thus highlighting the model's ability to detect all pertinent instances. The F1-score is a balanced measure that calculates the harmonic mean of precision and recall, providing a comprehensive assessment of the model's performance. The formulas for precision, recall, and F1-score are outlined in Equations 1–3, where TP denotes true positives, FP denotes false positives, and FN denotes false negatives:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$\mathrm{F1-score} = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

3.5 Baseline experiments

For the embedding models, FastText ("cc.zh.300") (24), SBERT ("distiluse-base-multilingual-cased"), and OpenAI's "text- embedding-3-large" were used to examine their performance on the test dataset, which consisted of 965 web-based medical documents. Each document contained a disease label and a textual query. To facilitate visualization, t-distributed Stochastic Neighbor Embedding(25) was initially applied to reduce the dimensionality of the embeddings. The classification performance of these models was then assessed using precision, recall, and F1-score as evaluation metrics.

For the LLMs comparison, GPT-4o, GPT-4o-mini, and Mistral were employed, focusing on their precision, recall, and F1-score on the test dataset. Furthermore, to evaluate the rationality of physician recommendations generated by these LLMs, a questionnaire was administered to three staff members involved in triaging. The participants were asked, "Based on your area of expertise, how would you rate the match between physician and the query?" Responses were measured using a 5-point Likert scale (26), with scores ranging from 1 (very inappropriate) to 5 (very appropriate). This evaluation did not involve relabeling the test dataset, but was used to assess whether the model's predictions were consistent with the professional judgment of these human experts in a triage scenario. The Mann–Whitney U test(27) was employed to determine if these assessments revealed any statistically significant differences.

4. Results

This study investigates the potential of applying RAG to improve the accuracy, reliability, and contextual relevance of physician recommendations. Similarly, the self-reflective retrieval-augmented framework proposed by Zhang et al. (2024) integrated a self-reflective correction mechanism, pharmacological memory bank, and RAG retrieval module, achieving an accuracy of 92.3% in pharmacological recommendation tasks. This confirms the potential of RAG technology to enhance recommendation reliability in specialized medical scenarios (Zhang et al., 2024). The objectives include analyzing the limitations of existing web-based medical services, evaluating the effectiveness of RAG in this context, and developing a framework for its implementation to enhance patient-physician matching. Ultimately, the research aims to provide insights that could markedly advance the personalization and efficacy of web-based medical services, thereby improving patient satisfaction. The study addresses the following research questions: How can RAG be utilized to effectively integrate extensive medical documents to provide personalized healthcare for patients? What impact do large language models and embedding models have on the quality of personalized recommendations? How effective is the RAG system in the context of medical services, particularly concerning the precision, recall, and F1-score of recommendations?

4.1 Data set summary

The dataset consists of 646,383 consultation records involving 193,675 patients and 858 physicians across 44 departments. According to Table 1, male patients constitute 32.95% (n = 212,983) of the records, while female patients make up 67.05% (n = 433,400). The age group most represented among patients is 20 to 39 years, accounting for 54.6% (n = 352,907) of the total consultations. Notably, senior physicians handled the majority of consultations, with 62.65% (n = 404,958) attributed to them. Additionally, the majority of response times were recorded at less than 90 min, comprising 38.13% (n = 246,472) of the total.

4.2 Evaluating the performance of embedding models

Figure 3 presents a comparative analysis of the clustering performance of three distinct embedding models using the test dataset. The models FastText, SBERT, and OpenAI were specifically developed for the purpose of labeling medical conditions. The evaluation focuses on the models' efficacy in differentiatingconditions such as cervical spondylosis, urinary tract infection, allergy, and diabetes, which are represented with distinct color- coded labels.

The initial plot illustrates the moderate capacity of the FastText model to differentiate between a numbers of medical conditions. Although the model is successful in creating clusters of similar labels, there is considerable overlap, indicating that there are challenges in effectively separating data points with identical labels. In contrast, the SBERT model demonstrates enhanced clustering capabilities, achieving a more distinct separation among different medical condition labels. This enhancement suggests SBERT's increased proficiency in distinguishing between conditions. The final plot reveals the performance of the OpenAI model, which exhibits the most distinct clustering. It forms well- defined, tightly grouped clusters corresponding to individual medical conditions and shows minimal overlap between different labels.

Table 2 provides a comparative analysis of the classification performance of three distinct embedding models: FastText, SBERT, and OpenAI, specifically in the context of medical condition labels using the test dataset. The analysis employs a classification model to predict disease labels from the embeddings of disease description texts. For each model, the precision, recall, and F1-score metrics are presented. The FastText model yielded a precision of 0.52, a recall of 0.44, and an F1-score of 0.46. In contrast, SBERT and OpenAI exhibited markedly superior performance, with both attaining high precision (0.95 and 0.97, respectively), recall (0.95 and 0.96, respectively), and F1-scores (0.95 and 0.96, respectively). These findings suggest that SBERT and OpenAI are more efficacious in accurately classifying medical condition labels from textual descriptions than FastText. In light of these findings and additional considerations, such as affordability, data security, and ease of migration, the study ultimately determined that the SBERT was the optimal embedding model.

TABLE 1 Summary of the characteristics of the collected data records (N = 646,383).

Characteristic		Value, *n* (%)
Gender		
	from Male	212,983 (32.95)
	from Female	433,400 (67.05)
Age of patient at consultation (years)		
	<20	118,484 (18.33)
	20–39	352,907 (54.6)
	40–59	125,957 (19.49)
	>60	49,035 (7.58)
Physicians’ professional title		
	Junior	10,766 (1.67)
	Intermediate	45,892 (7.1)
	Subsenior	184,767 (28.58)
	Senior	404,958 (62.65)
Physicians’ response time (minutes)		
	<90	246,472 (38.13)
	91–180	65,732 (10.17)
	181–270	41,033 (6.35)
	271–360	28,845 (4.46)
	>360	126,943 (19.64)
	Not response	137,358 (21.25)

4.3 Evaluating the performance of large language models (LLMs)

Table 3 presents a comparative analysis of three LLMs—Mistral, GPT-4o-mini, and GPT-4o—with a focus on their performance in terms of precision, recall, and F1-score in relation to physician recommendations using the test dataset. The Mistral model demonstrated a precision of 0.95, a recall of 0.94, and an F1-score of 0.94, indicating a balanced and efficient performance across all metrics. The GPT-4o-mini exhibited a precision of 0.95, which was comparable to that of Mistral. However, it demonstrated a slightly lower recall of 0.90 and consequently a reduced F1-score of 0.92. In contrast, the GPT-4o model exhibited a slightly lower precision (0.94) but a higher recall (0.97), resulting in the highest F1-score (0.95) among the models analyzed. Overall, the GPT-4o model demonstrated superior performance in synthesizing precision and recall, as reflected in its F1-score.

4.4 Rationality evaluation of LLMs

Table 4 presents a Mann–Whitney U test conducted on three pairs of LLMs using the test dataset. The Mistral model has been assigned a rating of 4.56, while the GPT-4o-mini has been rated 4.45 and the GPT-4o has been rated 4.67. The comparison between the Mistral and the GPT-4o- mini yielded a p-value of 0.003, indicating a statistically significant difference. The p-value for the comparison between Mistral and GPT-4o is 0.01, indicating a notable difference. Furthermore, the comparison between GPT-4o-mini and GPT-4o yielded a p-value of 0.001, thereby affirming the statistical significance of the difference. Considering the study's findings and additional factors such as affordability, data security, and ease of migration, the research ultimately concluded that the Mistral was the most suitable choice for implementation.

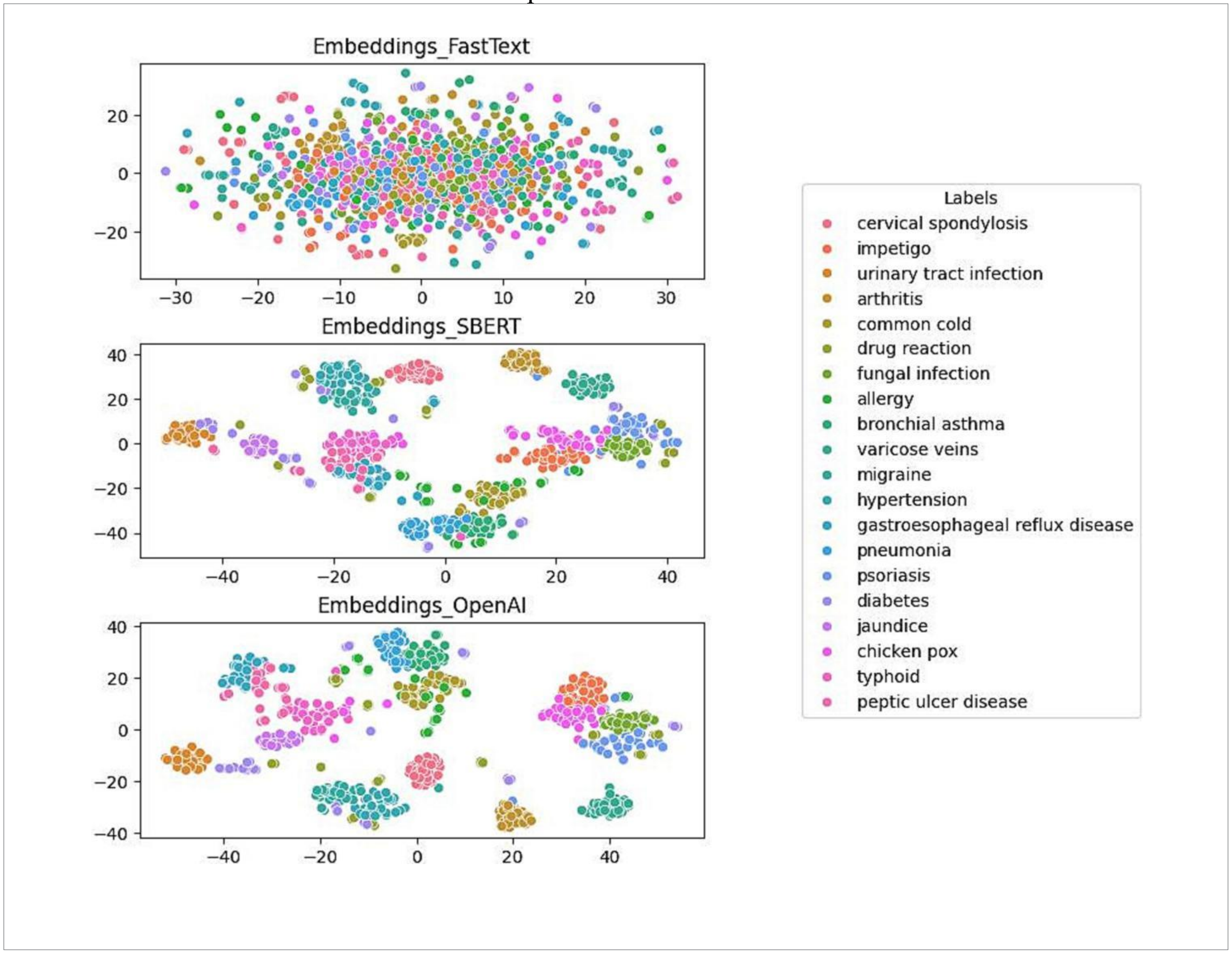


TABLE 2 Classification performance of FastText, SBERT, and OpenAI.

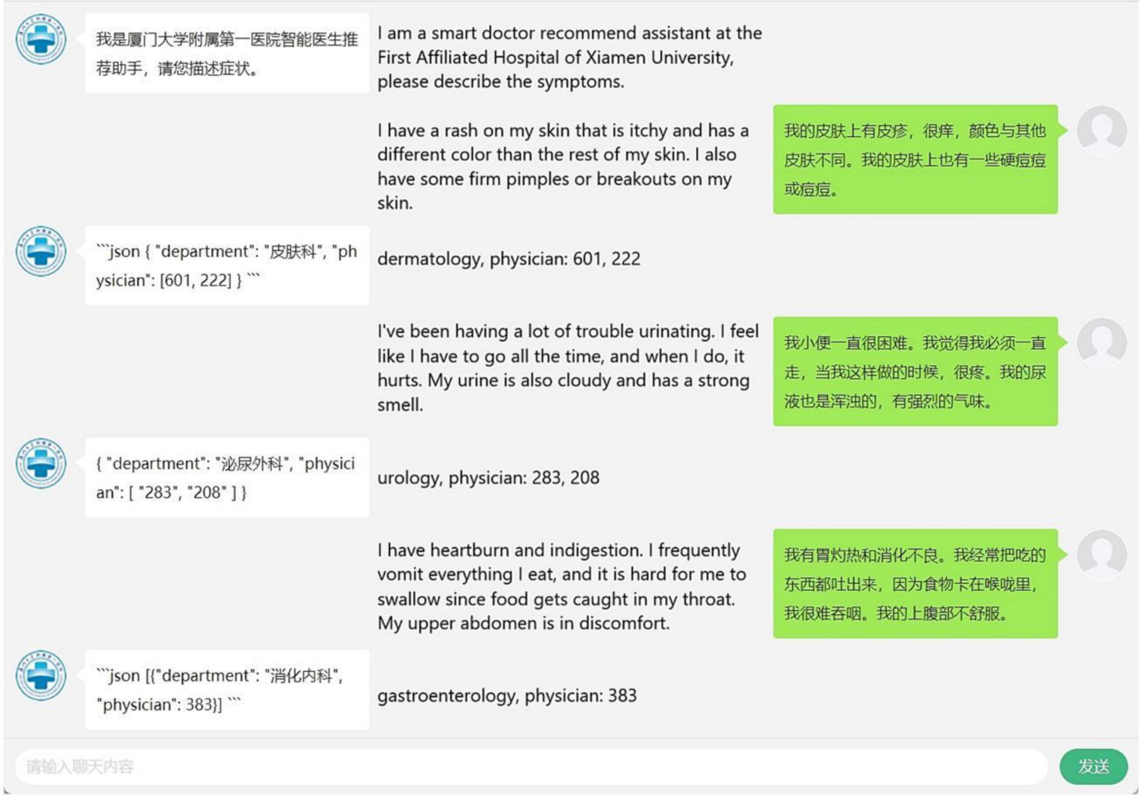
Embedding models	Precision	Recall	F_1-score
FastText	0.52	0.44	0.46
SBERT	0.95	0.95	0.95
OpenAI	0.97	0.96	0.96

TABLE 3 Comparative analysis of Mistral, GPT-4o-mini and GPT-4o.

LLMs	Precision	Recall	F_1-score
Mistral	0.95	0.94	0.94
GPT-4o-mini	0.95	0.90	0.92
GPT-4o	0.94	0.97	0.95

TABLE 4 The Mann–Whitney U test conducted on three pairs of LLMs.

Comparison	p value
Mistral vs. GPT-4o-mini	0.003
Mistral vs. GPT-4o	0.01
GPT-4o-mini vs. GPT-4o	0.001



我是厦门大学附属第一医院智能医生推荐助手，请您描述症状。

I am a smart doctor recommend assistant at the First Affiliated Hospital of Xiamen University, please describe the symptoms.

I have a rash on my skin that is itchy and has a different color than the rest of my skin. I also have some firm pimples or breakouts on my skin.

我的皮肤上有皮疹，很痒，颜色与其他皮肤不同。我的皮肤上也有一些硬痘痘或痘痘。

```json { "department": "皮肤科", "physician": [601, 222] } ```

dermatology, physician: 601, 222

I've been having a lot of trouble urinating. I feel like I have to go all the time, and when I do, it hurts. My urine is also cloudy and has a strong smell.

我小便一直很困难。我觉得我必须一直走，当我这样做的时候，很疼。我的尿液也是浑浊的，有强烈的气味。

{ "department": "泌尿外科", "physician": ["283", "208"] }

urology, physician: 283, 208

I have heartburn and indigestion. I frequently vomit everything I eat, and it is hard for me to swallow since food gets caught in my throat. My upper abdomen is in discomfort.

我有胃灼热和消化不良。我经常把吃的东西都吐出来，因为食物卡在喉咙里，我很难吞咽。我的上腹部不舒服。

```json [{"department": "消化内科", "physician": 383}] ```

gastroenterology, physician: 383

FIGURE 4

Performance of RAGPR model in mapping symptoms to medical departments and physicians. The layout of the image is divided into three sections: on the right are the human queries, on the left are the responses generated by the model, and in the middle are the interpretations of these interactions.

4.5 Case study

As illustrated in Figure 4, the RAGPR model demonstrates the capability to accurately identify medical specialties and recommend corresponding medical departments and physicians, based on a randomly selected set of symptom descriptions from the test database. For instance, the model correctly associates dermatological symptoms with the dermatology department and gastrointestinal symptoms with the gastroenterology department. This demonstrates that the RAGPR model effectively processes natural language descriptions to provide relevant medical recommendations, highlighting its potential for application in medical diagnostic systems.

5. Discussion

5.1 Principal findings

This study introduces an innovative physician triage algorithm called the RAGPR model, designed to enhance the accuracy and efficiency of web-based medical consultations. In our assessment, we evaluated various embedding and large language models to determine the most suitable options based on criteria such as cost- effectiveness, data security, and ease of migration. Consequently, the SBERT and Mistral models were selected as the optimal choices. The RAGPR model demonstrates an improved ability to accurately match patients' queries with physicians' specialties.

5.3 Reasons behind the performance achieved by the three algorithms

The performance of FastText, SBERT, and OpenAI's embedding models in the context of medical condition classification is reflective of their respective architectures and capabilities. FastText, although useful for capturing word representations, showed moderate clustering performance with noticeable overlap among medical conditions. This is primarily due to its focus on word-level embeddings without accounting for sentence-level semantics, limiting its effectiveness in distinguishing nuanced medical terms.

SBERT, on the other hand, provided a substantial performance boost. Its architecture, designed to derive sentence-level embeddings, allowed for more refined semantic understanding, resulting in distinct clustering for different medical labels. The ability to capture the contextual meaning of sentences led to significantly higher precision,recall, and F1-scores, making it highly effective for classifying medical conditions based on textual descriptions.

OpenAI's embedding model exhibited the most pronounced clustering capabilities, indicating its superior understanding and representation of semantic content. Its advanced architecture, likely with larger training datasets and refined algorithms, contributed to tightly grouped clusters and minimal label overlap. However, when considering additional factors such as cost, SBERT was identified as the optimal choice for the application, balancing high performance with practical implementation advantages.

5.4 Feasibility and potential extensions of the proposed model

The successful implementation of the RAGPR model in healthcare is contingent upon the existence of a robust IT infrastructure that is capable of handling large volumes of data, facilitating real-time processing, and integrating seamlessly with existing systems such as electronic health records (28). This may require the upgrading of existing systems or the adoption of cloud-based solutions (29) that offer scalability and flexibility. For the model to be widely adopted, it is essential that healthcare professionals receive adequate training to ensure effective use and the development of trust in its capabilities. Such training should include an understanding of how the model generates recommendations and the interpretation of its results. Furthermore, a comprehensive cost–benefit analysis is essential to assess whether the long-term benefits, such as enhanced efficiency, accuracy, and patient outcomes, justify the initial investment. Clearly articulated value propositions, such as reducing diagnostic errors, optimizing physician workload, or improving patient satisfaction, are vital for garnering stakeholder support.

One potential avenue for advancement within the field of healthcare is the integration of the RAGPR model into real-time decision support systems (30). Such systems could provide real-time recommendations during patient consultations, alerting physicians to potential problems such as drug interactions or abnormal test results, and suggesting next steps based on the latest clinical guidelines. This not only increases the efficiency of the visit, but also improves the quality of care by providing timely, evidence-based support. While current models focus on physician recommendations, future research could explore expanding these models to other areas, such as surgical recommendations, chronic disease management, or mental health support. Furthermore, the 'pharmacological memory bank' constructed by Zhang et al. (2024) demonstrates that accumulating historical recommendation cases and error-correction records can enhance the model's long-term reasoning ability. This mechanism can be migrated to the RAGPR model to optimize recommendation efficiency in repeated consultation scenarios and reduce matching errors for similar cases (Zhang et al., 2024). In addition, models could be tailored to specialty areas, such as oncology, cardiology, or pediatrics, to support complex decision- making processes. Another promising extension is the development of personalized medicine frameworks using predictive analytics. By analyzing patient-specific data over time, the model could predict future health risks, recommend preventive measures, and tailor treatment plans to a patient's unique health profile. This shift from reactive to proactive healthcare could significantly improve long-term patient outcomes.

5.5 Real-world application challenges

The implementation of the RAGPR model in healthcare is confronted with a multitude of considerable challenges, one of the most pivotal being the protection of patient data. The utilization of such confidential data for training and deployment must adhere to rigorous privacy regulations, such as HIPAA (31) in the United States and GDPR (32) in the European Union. Breaches of patient confidentiality can have profound legal and ethical ramifications. It is, therefore, of paramount importance to implement robust data encryption, anonymisation techniques, and secure data handling protocols to maintain patient trust and regulatory compliance.

A further significant challenge is the integration of the RAGPR model with existing healthcare IT infrastructure, such as electronic health records systems. This process can be hindered by a number of factors, including compatibility issues, inconsistencies in data format, and concerns regarding interoperability. It is not uncommon for the various systems utilized by healthcare institutions to lack the capacity to interact seamlessly with the model, necessitating substantial customization and development efforts.

Another challenge is the potential for bias, particularly when trained on datasets that are not representative of the population under study. Such bias has the potential to perpetuate or even exacerbate existing disparities in healthcare access and outcomes. For example, if the training datasets predominantly reflect certain demographic groups, the model's performance for underrepresented populations may be compromised, which could result in unequal treatment. To mitigate these biases, it is essential to employ diverse and representative training datasets and to incorporate fairness constraints during model development. This ensures that healthcare solutions are equitable.

Furthermore, the RAGPR model must undergo continuous adaptation in order to remain relevant and accurate in the context of evolving medical knowledge and practice. This necessitates the implementation of continuous learning frameworks that permit the model to update its knowledge base in response to new medical evidence, guidelines, and emerging diseases. Continuous monitoring, retraining, and validation mechanisms are essential to guarantee that the model provides recommendations that are up-to-date and reliable.

5.6 Limitations and future directions

The study has several important limitations. Firstly, the study was conducted exclusively within a single hospital, potentially limiting the applicability of the results to other settings or populations. Secondly, the dataset included irrelevant questions, such as "Doctor, will you be available tomorrow? Where can I find you?" These questions could introduce bias into the analysis. Lastly, a significant limitation of deep neural networks is their opacity, which refers to their lack of transparency in providing explanations for predictive results. This opacity poses challenges in understanding the rationale behind the predictions for specific samples.

Future research should seek to address the current limitations and explore potential avenues for improvement in patient-physician matching systems. A crucial objective is the development of sophisticated algorithms that enhance both the precision and responsiveness of this matching process. The incorporation of real- time data in conjunction with advanced machine learning models may facilitate the dynamic allocation of consultations based on physician availability, which could potentially reduce wait times and enhance patient satisfaction. Moreover, future studies should investigate the integration of multimodal data sources, including patient histories, imaging data, and real-time physiological signals. Such integration could facilitate a more comprehensive understanding of patient conditions, thereby improving diagnostic accuracy and treatment recommendations. Additionally, there is a need to develop explainable models that not only provide accurate recommendations but also offer transparent justifications for their decisions. This transparency would facilitate more informed decision-making in clinical settings, enhancing trust and effectiveness in healthcare.

6. Conclusion

This paper presents the RAGPR model, which is designed to improve the performance of triage in web-based medical services. The primary function of this model is to efficiently filter and select appropriate physicians, thereby assisting patients in identifying medical professionals best suited to address their specific healthcare needs. The implementation of this method has significant practical implications, suggesting its potential integration into various healthcare website systems to enhance the quality of physician recommendations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study complied with all relevant ethical regulations. All the available data sets have been deidentified and anonymized. The First Affiliated Hospital of Xiamen University Ethics Committee approved this study (approval number XMYY-2024KYSB400), and no informed consent was necessary.

Author contributions

YZ: Software, Visualization, Writing – original draft, Formal analysis, Methodology. YY: Formal analysis, Methodology, Writing – review & editing, Data curation, Resources, Software, Visualization.

SC: Data curation, Formal analysis, Project administration, Supervision, Validation, Visualization, Writing – original draft. YC: Conceptualization, Formal analysis, Methodology, Project administration, Software, Supervision, Writing – review & editing. KR: Formal analysis, Methodology, Validation, Writing – review & editing, Funding acquisition, Investigation, Project administration. YL: Data curation, Supervision, Formal analysis, Methodology, Validation, Visualization, Writing – review & editing. JZ: Funding acquisition, Investigation, Resources, Supervision, Writing – original draft, Data curation, Project administration. MZ: Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was funded by 2024 Xiamen Medical and Healthcare

Guidance Project “Research on Intelligent Diagnostic Model of Endometrial Pathology Images Based on Convolutional Neural Network Driven” (3502Z20214ZD1003).

Acknowledgments
The authors would like to thank all the participants in this study.

Conflict of interest
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note
All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Nordstrand AE, Anyan F, Bøe H, Hjemdal O, Noll L, Reichelt J, et al. Problematic anger among military personnel after combat deployment: prevalence and risk factors. *BMC Psychol*. (2024) 12:451. doi: 10.1186/s40359-024-01955-8

2. Sumar K, Blue L, Fatahi G, Sumar M, Alvarez S, Cons P, et al. The effect of adding physician recommendation in digitally-enabled outreach for COVID-19 vaccination in socially/economically disadvantaged populations. *BMC Public Health*. (2024) 24:1933. doi: 10.1186/s12889-024-18648-x

3. Brindisino F, Girardi G, Crestani M, et al. Rehabilitation in subjects with frozen shoulder: a survey of current (2023) clinical practice of Italian physiotherapists. *BMC Musculoskelet Disord*. (2024) 25:573. doi: 10.1186/s12891-024-07682-w

4. Rui JR, Guo J, Yang K. How do provider communication strategies predict online patient satisfaction? A content analysis of online patient-provider communication transcripts. *Digit Health*. (2024) 10:20552076241255617. doi: 10.1177/20552076241255617

5. Wetzel AJ, Koch R, Koch N, Klemmt M, Müller R, Preiser C, et al. 'Better see a doctor?' status quo of symptom checker apps in Germany: a cross-sectional survey with a mixed-methods design (CHECK.APP). *Health*. (2024) 10:20552076241231555. doi: 10.1177/20552076241231555

6. Iranzad R, Liu X, Dese K, Alkhadrawi H, Snoderly H, Bennewitz M. Structured adaptive boosting trees for detection of multicellular aggregates in fluorescence intravital microscopy. *Microvasc Res*. (2024) 156:104732:104732. doi: 10.1016/j.mvr.2024.104732

7. Herr K, Lu P, Diamreyan K, Xu H, Mendonca E, Weaver KN, et al. Estimating prevalence of rare genetic disease diagnoses using electronic health records in a children's hospital. *HGG Adv*. (2024) 5:100341. doi: 10.1016/j.xhgg.2024.100341

8. Lilli L, Bosello SL, Antenucci L, Patarnello S, Ortolan A, Lenkowicz J, et al. A comprehensive natural language processing pipeline for the chronic lupus disease. *Stud Health Technol Inform*. (2024) 316:909–13. doi: 10.3233/SHTI240559

9. Bonomo M, Rombo SE. Neighborhood based computational approaches for the prediction of lncRNA-disease associations. *BMC Bioinformatics*. (2024) 25:187. doi: 10.1186/s12859-024-

05777-8

10. Chew LJ, Haw SC, Subramaniam S. A hybrid recommender system based on data enrichment on the ontology modelling. *F1000Res*. (2021) 10:937. doi: 10.12688/ f1000research.73060.1

11. Abdullahi T, Mercurio L, Singh R, Eickhoff C. Retrieval-based diagnostic decision support: mixed methods study. *JMIR Med Inform*. (2024) 12:e50209. doi: 10.2196/50209

12. Yazaki M, Maki S, Furuya T, Inoue K, Nagai K, Nagashima Y, et al. Emergency patient triage improvement through a retrieval-augmented generation enhanced large- scale language model. *Prehosp Emerg Care*. (2024) 400:1–7. doi: 10.1080/10903127.2024.2374400

13. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi S, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. *Asian J Psychiatr*. (2024) 100:104168. doi: 10.1016/j.ajp.2024.104168

14. Arun G, Perumal V, Urias F, Ler Y, Tan B, Vallabhajosyula R, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: a comparative pilot study. *Anat Sci Educ*. (2024) 17:1396–405. doi: 10.1002/ase.2502

15. Tabaie A, Tran A, Calabria T, Bennett S, Milicia A, Weintraub W, et al. Evaluation of a natural language processing approach to identify diagnostic errors and analysis of safety learning system case review data: retrospective cohort study. *J Med Internet Res.* (2024) 26:e50935. doi: 10.2196/50935

16. Sharif S, Ghouchan R, Abbassian H, Eslami S. Comparison of regression methods to predict the first spike latency in response to an external stimulus in intracellular recordings for cerebellar cells. *Stud Health Technol Inform*. (2024) 316:796–800. doi: 10.3233/SHTI240531

17. Santander-Cruz Y, Salazar-Colores S, Paredes-Garcia WJ, et al. Semantic feature extraction using SBERT for dementia detection. *Brain Sci*. (2022) 12:270. doi: 10.3390/ brainsci12020270

18. Izzidien A, Fitz S, Romero P, et al. Developing a sentence level fairness metric using word embeddings. *Int J Digit Humanit*. (2022) 10:1–36. doi: 10.1007/ s42803-022-00049-4

19. Oh J, Park H. Effects of changes in environmental color Chroma on heart rate variability and stress by gender. *Int J Environ Res Public Health*. (2022) 19:711. doi: 10.3390/ijerph19095711

20. Santana EFM, Araujo JE. Realistic Vue: a new three-dimensional surface rendering approach for the in utero visualization of embryos and fetuses. *Radiol Bras*. (2019) 52:172–3. doi: 10.1590/0100-3984.2018.0050

21. Jolley KA, Bray JE, Maiden MCJ. A RESTful application programming interface for the PubMLST molecular typing and genome databases. *Database*. (2017) 2017:60. doi: 10.1093/database/bax060

22. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med*. (2024) 7:16. doi: 10.1038/s41746-023-00989-3

23. Tai ICY, Wong ECK, Wu JT, et al. Exploring offiine large language models for clinical information extraction: a study of renal histopathological reports of lupus nephritis patients. *Stud Health Technol Inform*. (2024) 316:899–903. doi: 10.3233/ SHTI240557

24. Endalie D, Haile G, Taye W. Deep learning-based idiomatic expression recognition for the Amharic language. *PLoS One*. (2023) 18:e0295339. doi: 10.1371/journal. pone.0295339

25. Saito Y, Itakura K, Ohtake N, et al. Classification of soybean chemical characteristics by excitation emission matrix coupled with t-SNE dimensionality reduction. *Spectrochim Acta A Mol Biomol Spectrosc*. (2024) 322:124785. doi: 10.1016/j. saa.2024.124785

26. Clements F, Vedam H, Chung Y, et al. Patient preference of level I, II and III sleep diagnostic tests to diagnose obstructive sleep apnoea among pregnant women in early to mid-gestation. *Sleep Breath*. (2024) 28:2387–95. doi: 10.1007/s11325-024-03114-0

27. Shu D, Zou G. Sample size planning for estimating the global win probability with precision and assurance. *Contemp Clin Trials*. (2024) 146:107665. doi: 10.1016/j. cct.2024.107665

28. Muayad J, Loya A, Hussain ZS, Chauhan M, Alsoudi A, de T, et al. Comparative effects of glucagon-like peptide 1 receptor agonists and metformin on glaucoma risk in patients with type 2 diabetes. *Ophthalmology*. (2024) 23:S0161–6420. doi: 10.1016/j. ophtha.2024.08.023

29. Bertò G, Rooks LT, Broglio SP, McAllister T, McCrea M, Pasquina P, et al. Diffusion tensor analysis of white matter tracts is prognostic of persisting post-concussion symptoms in collegiate athletes. *Neuroimage Clin*. (2024) 43:103646:103646. doi: 10.1016/j.nicl.2024.103646

30. Pardo E, Le Cam E, Verdonk F. Artificial intelligence and nonoperating room anesthesia. *Curr Opin Anaesthesiol*. (2024) 37:413–20. doi: 10.1097/ ACO.0000000000001388

31. Gottardelli B, Gatta R, Nucciarelli L, Tudor A, Tavazzi E, Vallati M, et al. GEN- RWD sandbox: bridging the gap between hospital data privacy and external research insights with distributed analytics. *BMC Med Inform Decis Mak*. (2024) 24:170. doi: 10.1186/s12911-024-02549-5

32. Wyatt KD, Minard-Colin V, Schleiermacher G, Willi M, Volchenboum S. GDPR and data sharing: the pediatric Cancer data commons experience. *Lancet Oncol*. (2024) 25:e227. doi: 10.1016/S1470-2045(24)00250-X

33. Zhaoyan Zhang, Yu Qiao, & Peimin Lu. (2024). Self-Reflective Retrieval-Augmented Framework for Reliable Pharmacological Recommendations. *Journal of Computational Methods in Engineering Applications*, *4*(1), 1–12. https://doi.org/10.62836/jcmea.v4i1.040108