EasyChair Preprint

№ 15757

IPF-RMF: Intelligent Patient Follow-up Supported by RAG and Multi-Model Fusion

Junyu Jia, Yue Qian, Zhe Shen, Zhi Wang, Dacheng Sang, Lanxin Yang, Lin Xu and Fulong Chen

January 24, 2025

IPF-RMF: Intelligent Patient Follow-up Supported by RAG and Multi-Model Fusion

Junyu Jia[1], Yue Qian[1], Zhe Shen[1], Zhi Wang[1], Dacheng Sang[2], Lanxin Yang[3], Lin Xu[1], and Fulong Chen[1⋆]

[1] School of Computer and Information, Anhui Normal University, Wuhu, China
{jjyu,qianyue,wangzhi,linxu,long005}@ahnu.edu.cn, smwy0110@163.com
[2] Department of Orthopaedics, Peking University Third Hospital, Beijing, China
dachengsang1616@163.com
[3] State Key Laboratory of Novel Software Technology, Software Institute, Nanjing University, China
lxyang@nju.edu.cn

Abstract. In modern healthcare services, intelligent patient follow-up is a critical approach to improving the quality of medical services and the efficiency of patient health management. Our study proposes an intelligent patient follow-up method based on Retrieval Augmented Generation (RAG) and multi-model fusion. Patient data and case information are collected using custom forms, forming an RAG-supported retrieval database to store follow-up records and relevant details. A multi-model framework was designed, using machine learning algorithms to predict key information such as follow-up schedules, and leveraging multiple large language models to generate initial follow-up recommendations. A decision-making large language model was utilized to integrate the initial follow-up recommendations from various language models, optimizing and developing the final personalized follow-up plan. Manual assessments were conducted to comprehensively analyze the quality of the final follow-up plan in terms of readability, professionalism, and other dimensions to evaluate the proposed method. Experimental results demonstrate that the proposed method significantly enhances the scientific validity and personalization of the follow-up plan, providing a reliable technical foundation for intelligent health management.

Keywords: Intelligent Patient Follow-Up · Retrieval Augmented Generation · Multi-Model Fusion · LLM.

1 Introduction

In modern medical services, Intelligent patient follow-up is a critical approach to improving medical service quality and the efficiency of patient health management[34]. With the development of information technology, Electronic Health Record (EHR) systems and patient management platforms have been widely

⋆ Corresponding author: long005@ahnu.edu.cn

used[30, 27], greatly promoting the digitization and standardization of medical information[6]. However, despite significant progress, existing medical information systems still face many challenges.

Existing systems fall short in personalized care and intelligent follow-ups. Traditional EHR systems only record diagnostic results and treatment history, lacking integration of patient characteristics and lifestyles to provide tailored treatment recommendations[1, 22]. Meanwhile, follow-up tools are not intelligent enough, lacking automated data analysis to schedule or adjust follow-ups[19]. These issues hinder optimal care planning, increase workload, reduce patient engagement, and negatively impact treatment and long-term health management.

To address these challenges, this study proposes an Intelligent Patient Follow-up Supported by RAG and Multi-Model Fusion (IPF-RMF) method. RAG technology mitigates the deficiencies of traditional generative models in terms of knowledge accuracy and contextual integrity by incorporating external knowledge bases and real-time retrieval mechanisms[8]. Building on this, we designed a multi-model framework that leverages machine learning algorithms and large-scale language models to generate personalized treatment recommendations and dynamic follow-up plans, thereby enhancing the scientific and personalized nature of follow-up strategies.

To verify the effectiveness of the proposed method, we conducted a detailed experimental study. The final follow-up plan was manually evaluated from various dimensions such as readability and professionalism. The results show that the IPF-RMF method significantly enhances the scientific and personalized nature of the follow-up plans, providing a solid technical foundation for intelligent health management. Additionally, the method exhibits good scalability and adaptability, making it applicable to various medical scenarios and likely to become an important development direction in intelligent healthcare.

The structure of this article is organized as follows. The second part reviews related work in the field of intelligent patient follow-up, focusing on RAG and multi-model fusion techniques. The third part introduces the background of this study, providing a detailed explanation of the design and implementation of the IPF-RMF framework. The fourth part elaborates on the specific methods of the intelligent follow-up process, including data collection, model integration, and the generation of personalized follow-up plans. The fifth part evaluates the effectiveness of the proposed method through experiments and analyzes its performance in improving the scientificity and personalization of follow-up plans. Finally, the sixth section summarizes the main conclusions and suggests future research directions.

2 Related Work

Extensive research has been conducted in medical data management and patient follow-up, achieving significant results[4, 12]. Recently, RAG technology and multi-model fusion have gained widespread attention. RAG significantly improves medical data management by integrating knowledge retrieval and gen-

eration models[16]. For example, Liu et al.[23] proposed a fine-tuning method combining pre-trained parameter memory with neural retrieval, which enhanced model performance in knowledge-intensive NLP tasks, especially in open-domain QA. Studies indicate that RAG dynamically integrates patient history, the latest medical literature, and external knowledge to generate more precise and personalized treatment recommendations[32].

Multi-model fusion enhances data analysis and decision-making accuracy by integrating various machine learning models[13]. For instance, Zheng et al.[36] proposed a multimodal graph learning framework (MMGL) that captures intermodal correlations and achieves superior performance in disease prediction tasks. Additionally, Chen et al.[9] highlighted emerging trends in multimodal medical image processing, such as generative adversarial networks and contrastive learning. These studies demonstrate the advantages of multi-model fusion in providing comprehensive and accurate diagnostic and treatment recommendations.

Existing methods can be categorized into single-model and multi-model fusion approaches. Single-model methods typically rely on specific data sources or model types, such as traditional EHR systems, which utilize structured data and standardized forms for management[14, 15]. While simple to implement, these methods lack flexibility and personalization, and struggle with complex medical data. In contrast, multi-model fusion methods enhance overall performance by integrating diverse data sources and model types. For example, multimodal deep learning frameworks can simultaneously process text, images, and physiological signals[21]. However, they are more complex to implement and require addressing data source integration challenges[20, 3].

Despite the potential of RAG and multi-model fusion in medical data management and patient follow-up[24], limitations remain. Many systems rely on manual input, resulting in inefficient and error-prone data collection. Follow-up tools lack personalized reminder functions and fail to respond swiftly to patient needs. Furthermore, they often do not fully utilize medical research and external knowledge bases, leading to insufficient scientific and personalized plans. The rigid design of many systems hampers their adaptability to doctor-patient interactions, affecting user experience.

This study addresses the limitations of existing patient follow-up systems by introducing RAG technology and multi-model fusion. We propose an intelligent follow-up platform that integrates RAG technology, custom forms, and multi-model integrated prediction algorithms. This platform enhances flexibility, personalization, and data analysis capabilities while promoting efficient doctor-patient communication through real-time retrieval and generation functions, providing more accurate treatment recommendations and follow-up plans. Specifically, it combines the capabilities of knowledge retrieval and generation models to dynamically integrate patient history, the latest medical literature, and external knowledge for personalized follow-up suggestions. Leveraging multi-model fusion, the platform comprehensively analyzes various data sources to deliver accurate and holistic diagnostic and treatment advice. Custom forms enable efficient patient data collection and case information management, forming a

retrieval database that supports RAG technology and improves data collection quality and efficiency. This study not only fills existing research gaps but also offers a novel solution for intelligent follow-up, significantly enhancing the quality and efficiency of medical services[5].

3 The IPF-RMF Framework

3.1 IPF-RMF Architecture

Fig.1 illustrates the complete IPF-RMF architecture process, encompassing data collection, the integration of RAG and multi-model fusion, and the generation of personalized follow-ups. The architecture consists of the RAG module and the multi-model fusion module, designed to deliver precise and personalized healthcare support through efficient data management and processing.

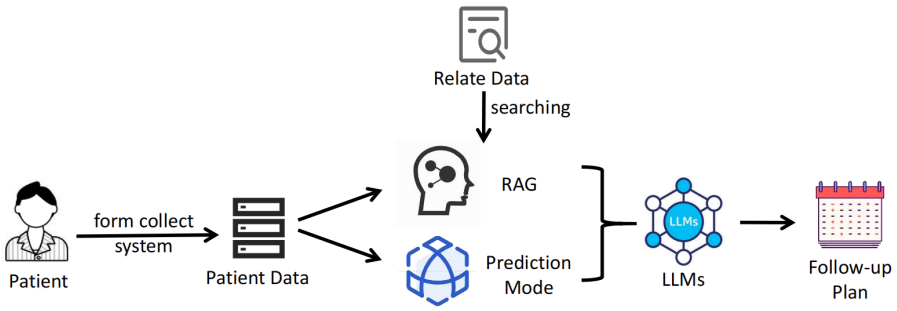

Fig. 1. IPF-RMF architecture

The RAG module focuses on retrieving and enhancing historical patient data and follow-up records. Custom forms efficiently collect medical data, which is stored in a retrieval-enhanced database integrating basic patient information, follow-up records, and treatment plans. RAG technology identifies similar case data, compensating for current data gaps and providing comprehensive input to enhance the adaptability and accuracy of treatment plans.

The multi-model fusion module integrates multiple Large Language Models(LLM) and Machine Learning(ML) algorithms to generate multi-dimensional follow-up recommendations. Leveraging the augmented data from the RAG module, models generate initial suggestions considering disease conditions, treatment responses, and individual needs. The decision model then optimizes these outputs, formulating the most suitable personalized follow-up plan.

The IPF-RMF architecture, which combines RAG and multi-model fusion technologies, provides an efficient and personalized follow-up solution that offers substantial support to clinical decision-making processes and enables more intelligent and personalized healthcare management. The IPF-RMF architecture

combines RAG and multi-model fusion technologies to deliver efficient and personalized follow-up solutions, supporting clinical decision-making and advancing intelligent healthcare management.

3.2 RAG Module

The RAG module relies on comprehensive medical data, collected via the methods illustrated in Fig.2. Retrieving symptoms and mathematical features from medical databases, clinical manifestations and treatment plans from hospital databases, and all existing medical data for specific patients via a custom form collection system. All data is persistently stored in JSON format. Before use, keyword extraction and text segmentation are performed. Segmentation is sentence-based to maintain semantic integrity. Short sentences are combined to approximate 256/512 tokens, enabling processing by LLMs.

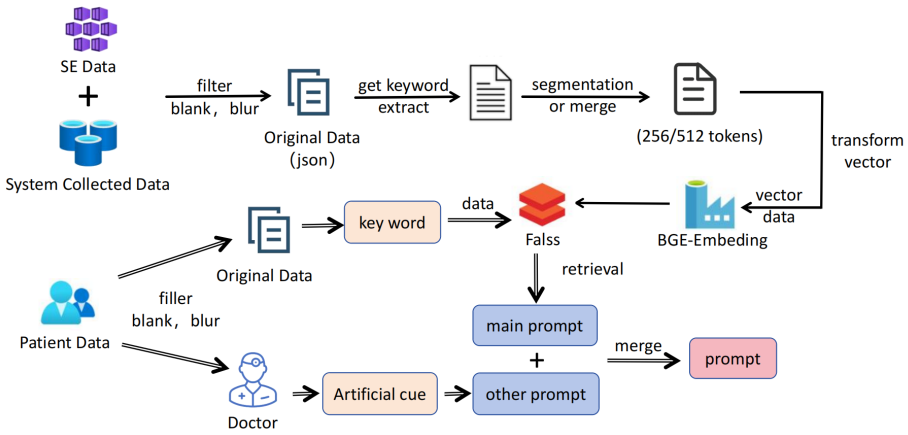

Fig. 2. RAG module

To meet RAG's efficiency requirements for data processing and retrieval, vectorization[2] was employed. Text data was vectorized using the BGE-Embedding model, with local fine-tuning improving the indexing accuracy for medical data by over 27%. The indexed data is then stored in a FAISS database, enabling efficient storage and retrieval.

Before IPF-RMF usage by medical practitioners, RAG retrieves relevant data from the database based on queries and patient information using efficient indexing methods, integrating it into the prompt. Similarity algorithms filter data exceeding a similarity threshold for prompt inclusion. The initial prompt is then enhanced by adding relevant information, increasing its specificity and logical coherence, thereby improving LLM understanding and output accuracy.

3.3 Multi-Model Fusion Module

The Multi-Model Fusion module leverages machine learning algorithms combined with medical data from the local knowledge base to predict key information in the follow-up plan. These algorithms account for individual patient differences, such as the severity of their condition and treatment response time, to create personalized follow-up plans. To facilitate subsequent processing and analysis, the module standardizes the knowledge base data and integrates the predicted follow-up plans with existing data, generating standardized statements containing medical causal logic and logical types. These statements are stored as key-value pairs for ease of future processing and querying.

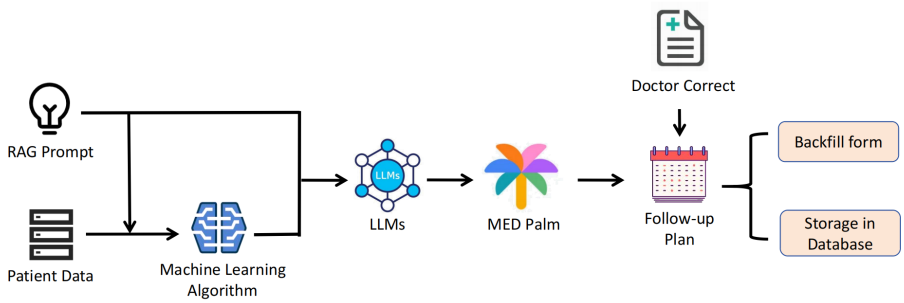

Fig. 3. Multi-Model Fusion module

As shown in Fig.3, standardized statements combined with prompts generated by RAG are provided as inputs to multiple large language models (LLMs) for further analysis[17]. These models comprehensively consider data from the knowledge base and personalized follow-up time predictions, generating multiple potential follow-up treatment plans for each patient, including future follow-up arrangements. Ultimately, a decision-making model integrates the outputs from these models with patient-specific details to propose the most suitable treatment plan. This process not only focuses on treatment effectiveness but also incorporates the patient's personal preferences and lifestyle.

After each follow-up, the system updates the patient's condition records and treatment progress based on the latest results. This dynamic update mechanism ensures that the follow-up plan can be flexibly adjusted to reflect changes in the patient's condition, while also providing critical reference information for future follow-ups. Additionally, the system interprets the outputs from the large model, assisting doctors in pre-filling sections of follow-up forms, such as recommendations for subsequent treatments or recovery arrangements. This significantly reduces doctors' workloads and enhances the efficiency and accuracy of the follow-up process. Doctors can refine and enhance the preliminary plans, ultimately creating personalized follow-up reports for each patient.

4 Framework Evaluation

4.1 Experimental Set Up

Dataset. To comprehensively evaluate the effectiveness of the IPF-RMF framework in intelligent patient follow-up plan generation, this study conducts a systematic comparison and analysis using the Medical Information Mart for Intensive Care IV dataset(MIMIC-IV). MIMIC-IV is a widely used public medical database developed collaboratively by the Massachusetts Institute of Technology's Computational Physiology Laboratory and the Beth Israel Deaconess Medical Center. The dataset contains anonymized electronic health records of patients treated in the intensive care units of Beth Israel Deaconess Medical Center from 2008 to 2019. MIMIC-IV provides detailed clinical information, including patient admission and discharge summaries, laboratory test results, medication prescriptions, surgical records, and more. In this study, we selected 1500 follow-up records with varying symptoms and severity levels, and based on different treatment plans over time, defined the time periods after the first follow-up as the follow-up treatment plans, which were used as a reference for subsequent experiments.We also adopted several key performance evaluation metrics to comprehensively measure the performance of different models or model combinations.

Framework Configuration. We selected four different model combinations for comparative experiments to explore how LLMs and their combinations can provide personalized follow-up recommendations based on the specific conditions of patients. Specifically, our research includes a case where only the base LLM model is used to test its basic ability to generate follow-up plans. A model that combines LLM with RAG to enhance the relevance and precision of the output by retrieving relevant historical data. A method that integrates results from multiple independent LLM models to combine diverse information sources and improve the overall quality of the plan. A method that employs both RAG technology and multi-model fusion strategies, aiming to achieve the best level of personalized service and precision.

To identify the most effective model combinations, we considered several different LLM configurations, including BERT[11], T5[26], XLNet [33] and GPT-4[31, 25]. Each combination is paired with a decision support model, Medpalm[28], to assist the final decision-making process.

We begin by conducting a comprehensive evaluation of the individual LLMs, followed by an analysis of various LLM combinations. In order to enhance the accuracy of follow-up plan generation while minimizing the risks of overfitting and excessive computational complexity[18], we selected a combination of four individual LLMs[35] to achieve the best effect improvement. The term "LLMs" encompasses BERT, T5, XLNet, and GPT-4.

To comprehensively evaluate the capabilities of various model configurations, we designed a series of prompts ranging from simple to complex for testing. These

prompts covered basic information queries (e.g., patient age, gender) to more in-depth questions (e.g., providing specific health management suggestions). This approach allows us to observe the performance differences between different models when faced with varying levels of challenges.

Evaluation. The evaluation process mainly relies on manual scoring, considering multiple dimensions such as accuracy, personalization, response time, and user experience.

Accuracy Scoring(AS). Scoring is based on the degree of match between the generated follow-up plan and the actual case. The scoring criteria include the accuracy of the treatment plan, the reasonableness of the follow-up schedule, etc[29].

Readability Scoring(RS). Evaluate whether the generated follow-up plan is easy to understand, whether the language is clear and concise, and whether it allows doctors to grasp the key points of the patient's condition quickly.

Professionalism Scoring(PS). Assess the medical standardization of the follow-up plan, whether it covers all aspects of the patient's condition (such as symptoms, medical history, treatment responses, etc.), and provides appropriate medical recommendations[7].

Response Time Scoring(RTS). Record the time taken by the model to generate the follow-up plan and score it based on the response speed[10].

To ensure the fairness and objectivity of the evaluation, we employ experts from the medical field to score the models. Specific scoring criteria are set for each dimension, and three experts review the output of each model. The average score from these experts is then taken as the final evaluation result, with the highest-scoring model being selected. The evaluation process is divided into two stages: initial scoring and final scoring.

Initial Scoring. Each expert gives an initial score for the output of the four models based on the preset dimensions. Each expert reviews 500 data samples and scores the outputs. Clarify and standardize the scoring criteria for each dimension before the process begins to ensure consistency in scoring.

Final Scoring. The expert scores for each model are aggregated to calculate the final score for each model. The score for each dimension is weighted based on its importance and averaged, and the model with the highest score is selected. Final Score(FS) is calculated as:

$$FS = w_1 \times AS + w_2 \times RS + w_3 \times PS + w_4 \times RTS \tag{1}$$

Where w_1, w_2, w_3, w_4 represent the weights for accuracy, readability, professionalism, and response time, respectively. After detailed discussions among the experts, the weights for each scoring dimension are determined as follows: $w_1 = 40\%$, $w_2 = 20\%$, $w_3 = 20\%$, and $w_4 = 20\%$.

By comprehensively considering the factors mentioned above and assigning different weights based on the importance of each metric. We aim to conduct a comprehensive and detailed evaluation of the performance of different model configurations under the IPF-RMF framework in the intelligent generation of

patient follow-up plans. This approach not only helps identify the most effective model combinations but also provides a scientific basis for further optimization of such systems in the future.

4.2 Results

Based on the experimental results, the RAG + Multiple LLM Fusion Model outperforms other models across all dimensions, especially in accuracy, readability, and professionalism. However, this model has a relatively longer response time and may require further optimization to enhance its real-time performance. Overall, this model demonstrates great potential in generating intelligent follow-up plans for patients, particularly in medical scenarios that require high levels of personalization and professionalism.

Results on simple prompts are denoted as AS1, RS1, PS1, RTS1, FS1, while results on complex prompts are denoted as AS2, RS2, PS2, RTS2, FS2. This setup allowed us to assess model performance across different levels of task complexity.

Table 1. Evaluation of Individual LLMs and LLM Combinations

Model	AS1	RS1	PS1	RTS1	FS1	AS2	RS2	PS2	RTS2	FS2
BERT	48	62	70	74	60.4	51	62	71	74	61.8
T5	51	64	73	77	63.2	53	64	72	75	63.4
XLNet	47	60	69	73	59.2	50	61	69	73	60.6
GPT-4	49	61	71	75	61.0	51	63	72	76	62.6
LLMs	**56**	**68**	**78**	**82**	**71.2**	**58**	**69**	**77**	**81**	**73.2**

Table 1 evaluates four individual LLM configurations (GPT-4, BERT, T5, RoBERTa), and shows the evaluation results of different LLM combinations. Each individual model performs differently across various metrics, highlighting the strengths and weaknesses of each in certain tasks. Compared to single LLMs, combination models generally perform better across various tasks. The combined LLMs show significant improvements in metrics such as AS1, RS1, and PS1, demonstrating greater adaptability and flexibility.

Table 2. Evaluation of Machine Learning + LLM & LLMs

Model	AS1	RS1	PS1	RTS1	FS1	AS2	RS2	PS2	RTS2	FS2
ML + BERT	56	70	79	80	68.2	58	70	78	81	69.0
ML + T5	58	71	80	**82**	69.8	60	72	**80**	**83**	71.0
ML + XLNet	55	69	78	79	67.2	57	69	77	81	68.2
ML + GPT-4	57	71	80	81	69.2	59	72	80	**83**	70.6
ML + LLMs	**59**	**72**	**81**	**82**	**70.6**	**61**	**73**	**80**	**83**	**71.6**

Table 2 evaluates the combination of machine learning models with LLMs and the effects of combining machine learning with multiple LLMs. The combined models further enhance performance across multiple tasks, especially in tasks such as PS1 and RTS1, showing that the combination of machine learning algorithms and LLMs provides additional advantages, improving accuracy and robustness. Compared to single LLMs and the machine learning + single LLM combinations, machine learning with multiple LLMs further improves performance in tasks such as AS1, RS1, etc., showing more balanced improvements across most metrics.

Table 3. Evaluation of RAG + LLM & LLMs

Model	AS1	RS1	PS1	RTS1	FS1	AS2	RS2	PS2	RTS2	FS2
RAG + BERT	58	70	77	78	68.2	60	71	79	81	70.2
RAG + T5	60	**72**	**80**	**82**	70.8	62	73	81	83	72.2
RAG + XLNet	57	69	78	80	68.2	59	71	79	82	70.0
RAG + GPT-4	59	71	**80**	**82**	70.2	61	72	80	84	71.6
RAG + LLMs	**64**	67	79	81	**71.0**	**65**	**76**	**82**	**85**	**74.6**

Table 3 evaluates the combination of RAG (retrieval-augmented generation models) and LLMs, and the evaluation results after combining RAG with multiple LLMs. The combination of RAG and LLMs shows good performance across multiple evaluation metrics, especially in tasks such as AS1, RS1, and PS1, where the combination outperforms the single models significantly. The combination of RAG and multiple LLMs performs excellently across all evaluation metrics, particularly in tasks such as AS1, RS1, RTS1, where it shows a significant advantage, indicating that this configuration effectively enhances the overall performance of the model.

Table 4. Evaluation of RAG + Machine Learning + LLM & LLMs

Model	AS1	RS1	PS1	RTS1	FS1	AS2	RS2	PS2	RTS2	FS2
RAG + ML + BERT	60	72	79	81	70.4	62	74	80	83	72.2
RAG + ML + T5	62	74	**81**	**83**	72.4	64	75	82	**85**	74.0
RAG + ML + XLNet	59	71	78	80	69.4	61	72	79	82	71.0
RAG + ML + CPT-4	61	73	79	81	71.0	63	74	81	84	73.0
RAG + ML + LLMs	**63**	**75**	**81**	**83**	**73.0**	**70**	**84**	**86**	77	**77.4**

Table 4 shows the performance of the combination of RAG, machine learning, and LLMs and shows the evaluation results of combining RAG, machine learning, and multiple LLMs. The combined models (such as RAG + ML + BERT to RAG + ML + LLMs) show significant performance improvements across multiple evaluation tasks, especially in metrics such as AS1, RS1, RTS1, and FS2, where the overall capability of the model is greatly enhanced. The combination of RAG

and machine learning with multiple LLMs performs the best, demonstrating the model's powerful performance and efficient personalization capability.

5 Discussion

In this study, we propose an intelligent patient follow-up plan generation method based on the IPF-RMF framework and compare four different model configurations when handling prompts of varying complexity. By evaluating accuracy, readability, professionalism, and response time, we demonstrate the advantages of this method in generating personalized and accurate follow-up plans.

The core advantage of the IPF-RMF framework lies in combining various models with the retrieval capabilities of RAG technology. By integrating historical case data into the generation process, RAG significantly enhances the personalization and accuracy of follow-up plans. Compared to traditional LLM models, incorporating RAG technology produces plans that better align with patients' actual conditions and treatment needs while improving medical compliance. This advantage is further amplified within the multi-model fusion framework.

Experimental results show that in terms of readability, RAG and multi-model fusion models outperform single LLM models. Models with RAG perform better, indicating that while multi-model fusion enriches content and perspectives, the absence of RAG historical data retrieval leads to deficiencies in language fluency and applicability.

Multi-model fusion results indicate that integrating outputs from different LLMs can enhance follow-up plan quality. By leveraging the characteristics of diverse language models, multi-LLM fusion provides more comprehensive and diversified information. However, a drawback is increased response time, especially with multiple large models involved. Although RAG shortens prompt generation time through retrieval enhancement, multi-model inputs and decision-making processes add extra response time.

To address response time issues, future work could focus on selecting lower-computation language models tailored for healthcare and introducing caching mechanisms for repeated prompts or similar cases to avoid full retrieval and reasoning every time. By caching model outputs and decision processes, results can be returned directly in similar scenarios, reducing response time. Additionally, pre-processing case data and prompts and precomputing generation paths can effectively shorten inference time.

6 Conclusion

This study introduces an intelligent patient follow-up method (IPF-RMF) based on RAG and multi-model fusion, designed to efficiently collect patient data through custom forms and generate personalized follow-up plans using advanced neural network techniques. Compared to traditional methods, this approach significantly enhances the scientific and personalized aspects of follow-up plans, especially in managing complex cases and diverse patient needs.

Compared to traditional follow-up methods, this study reveals that models leveraging RAG technology generate more accurate follow-up plans by retrieving historical data of similar cases. Multi-model fusion further boosts predictive capabilities, aligning recommendations more closely with individual patient needs, particularly in complex cases, demonstrating higher accuracy and reliability. Evaluations confirm that the combination of RAG and multi-model fusion excels in readability, professionalism, and personalized adaptation compared to single models.

Experimental results validated the practical application of the model through manual evaluation. When compared to actual treatment plans, the proposed method ensures treatment effectiveness while reducing doctors' workloads and improving doctor-patient communication efficiency. Data analysis and case studies confirm the feasibility and effectiveness of this approach in personalized healthcare and intelligent health management.

Despite its promising results, this study has certain limitations. Future work will focus on optimizing the model's generalization ability for complex clinical scenarios and improving accuracy and robustness using more real-world clinical data. Additionally, as AI technology evolves, the IPF-RMF method holds great potential for broader applications, particularly in chronic disease management, disease prediction, and personalized treatment plan generation.

References

1. Abul-Husn, N.S., Kenny, E.E.: Personalized medicine and the power of electronic health records. Cell **177**(1), 58–69 (2019)
2. Alkhalaf, M., Yu, P., Yin, M., et al.: Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. Journal of Biomedical Informatics **156**, 104662 (2024)
3. Amal, S., Safarnejad, L., Omiye, J., et al.: Use of multi-modal data and machine learning to improve cardiovascular disease care. Frontiers in Cardiovascular Medicine **9**, 840262 (2022)
4. Azbeg, K., Ouchetto, O., Jai Andaloussi, S.: Blockmedcare: A healthcare system based on iot, blockchain and ipfs for data management security. Egyptian Informatics Journal **23**(2), 329–343 (2022)
5. Bedi, S., Liu, Y., Orr-Ewing, L., et al.: Testing and evaluation of health care applications of large language models: A systematic review. JAMA (10 2024)
6. Bodin, O.N., Bezborodova, O.E., Mitroshin, A.N., et al.: Optimization of medical care provision in intelligent medical information system. In: 2023 Systems and Technologies of the Digital HealthCare (STDH). pp. 149–152 (2023)
7. Chang, Y., Wang, X., Wang, J., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (2024)
8. Chen, S., Guevara, M., Moningi, S., et al.: The effect of using a large language model to respond to patient messages. The Lancet Digital Health **6**(6), e379–e381 (2024)
9. Chen, X., Xie, H., Tao, X., et al.: Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. Artificial Intelligence Review **57**, 91 (2024)

10. Chen, Y., Qian, S., Tang, H., et al.: Longlora: Efficient fine-tuning of long-context large language models. arXiv preprint arXiv:2309.12307 (2023)
11. Devlin, J., Chang, M.W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
12. Dow, R., Chen, K.M., Zhao, Cindy S., et al.: Artificial intelligence improves patient follow-up in a diabetic retinopathy screening program. Clinical Ophthalmology **17**, 3323–3330 (2023)
13. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov: A guide to deep learning in healthcare. Nature medicine **25**(1), 24–29 (2019)
14. Gamal, A., Barakat, S., Rezk, A.: Standardized electronic health record data modeling and persistence: A comparative review. Journal of Biomedical Informatics **114**, 103670 (2021)
15. Heumos, L., Ehmele, P., Treis, e.a.: An open-source framework for end-to-end analysis of electronic health record data. Nature Medicine **30**, 3369–3380 (2024)
16. Jeong, M., Sohn, J., Sung, M., et al.: Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics **40**(1), i119–i129 (2024)
17. Ji, Y., Li, Z., Meng, R., et al.: Rag-rlrc-laysum at biolaysumm: Integrating retrieval-augmented generation and readability control for layman summarization of biomedical texts. In: BioNLP@ACL. pp. 810–817 (2024)
18. Khan, F., Kim, Y.S.: A novel hybrid ensemble model for classification of imbalanced datasets. Journal of King Saud University-Computer and Information Sciences (2020)
19. Klumpp, M., Hintze, M., Immonen, e.a.: Artificial intelligence for hospital health care: Application cases and answers to challenges in european hospitals. Healthcare **9**(8) (2021)
20. Krones, F., et al.: Review of multimodal machine learning approaches in healthcare (2024)
21. Li, K., Chen, C., Cao, W., et al.: Deaf: A multimodal deep learning framework for disease prediction. Computational Biology and Medicine **156**, 106715 (2023)
22. Li, Y.H., Li, Y.L., Wei, M.Y., et al.: Innovation and challenges of artificial intelligence technology in personalized healthcare. Scientific Reports **14**(1), 18994 (2024)
23. Liu, Y., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)
24. Meskó, B.: Prompt engineering as an important emerging skill for medical professionals: Tutorial. Journal of Medical Internet Research **25**, e50638 (2023)
25. OpenAI, Achiam, J., Adler, S., et al.: Gpt-4 technical report (2024)
26. Raffel, C., Shazeer, N., Roberts, A., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 1–67 (2020)
27. Raina, R., Jha, R.K.: Intelligent and interactive healthcare system (i2hs) using machine learning. IEEE Access **10**, 116402–116424 (2022)
28. Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. Nature **620**, 172–180 (2023)
29. Svetozarević, M., Janković, I., Lukić, S., et al.: Standards for use of llm in medical diagnosis. In: Information Society 2024. pp. 1–4 (2024)
30. Vardhan, M., Nathani, D., Vardhan, S., et al.: Large language models as synthetic electronic health record data generators. In: 2024 IEEE Conference on Artificial Intelligence (CAI). pp. 804–810 (2024)

31. Waisberg, E., Ong, J., Masalkhi, M., et al.: Gpt-4: A new era of artificial intelligence in medicine. Irish Journal of Medical Science (1971 -) **192**(6), 3197–3200 (2023)
32. Xiong, G., et al.: Benchmarking retrieval-augmented generation for medicine (2024)
33. Yang, Z., Dai, Z., Yang, Y., et al.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Neural Information Processing Systems (2019)
34. Zhang, F.: Design and implementation of a hospital patient follow-up system based on deep learning. In: 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC). pp. 216–220 (2023)
35. Zhang, X., Li, Y., Li, H.: Multimodal ensemble learning for medical diagnosis using hybrid deep learning models. Computers in Biology and Medicine **158**, 106596 (2023)
36. Zheng, S., Zhu, Z., Liu, Z., et al.: Multi-modal graph learning for disease prediction. IEEE Transactions on Medical Imaging **41**(9), 2207–2216 (2022)