ISSN 2959-6157

The Investigation of the Application of RAG Technology in the Field of EHR

Yifeng Xu

Department of Civil Engineering, Tongji University, Shanghai, China
Corresponding author: halyang@tongji.edu.cn

Abstract:

Electronic Health Records (EHR) play a crucial role in contemporary medical information systems, capturing extensive data in multiple scenarios. This review explores the impact of Retrieval Augmented Generation (RAG) on managing EHR. By integrating retrieval and generation processes, RAG significantly enhances data handling, clinical decision support, and patient management. EHRs contain extensive data like patient histories and diagnostic information, which are becoming increasingly complex. The RAG improves the accuracy and richness of this data processing by using a retrieval module to extract relevant text fragments and a generation module to create precise outputs. Despite its potential, RAG faces challenges such as data inconsistency, privacy concerns, and the need for efficient training. Recent studies highlight the RAG's effectiveness in summarizing clinical notes and enhancing prediction accuracy, suggesting a promising future in healthcare. However, ongoing research is necessary to optimize the RAG's application and address these challenges, aiming to transform healthcare delivery effectively.

Keywords: Retrieval augmented generation; electronic health records; large-scale language model.

1. Introduction

Electronic Health Records (EHR) are an important component of modern medical information systems, recording a large amount of data such as patient medical history, diagnostic information, treatment plans, and laboratory test results. With the continuous advancement of medical informatization, the amount and complexity of EHR data are also rapidly increasing. This not only provides abundant resources for medical research and practice, but also brings enormous challenges in data processing and information extraction.

In recent years, deep learning based Natural Language Processing (NLP) technology has shown great potential in EHR data analysis. Among them, the Retrieval Augmented Generation (RAG) model, as an emerging NLP technology, can effectively improve the accuracy and richness of information extraction by combining retrieval and generation processes. In the RAG model, the retrieval module is first used to retrieve text fragments related to the query from a large document library, and then the generation module is used to generate the final answer or text. This method not only utilizes the rich information of existing documents, but also provides coherent and accurate output through the generation module.

In the field of EHR, the application prospects of RAG models are broad. Its potential applications include but are not limited to the following aspects: 1) Clinical decision support: By integrating and analyzing a large amount of EHR data, the RAG model can provide targeted decision support suggestions for clinical doctors, improving the accuracy and efficiency of diagnosis and treatment. 2) Medical knowledge discovery: The RAG model can mine potential medical knowledge from EHR, reveal the potential associations and treatment effects of diseases, and provide new perspectives for medical research. 3) Patient management: By analyzing EHR data, the RAG model can help medical institutions better manage patients, provide personalized medical plans, and improve the level of patient health management. 4) Natural language queries: The RAG model can handle complex natural language queries, extract relevant information from EHR, and provide convenient information retrieval services for doctors and researchers.

Although the RAG model has shown great potential in the field of EHR, its application still faces many challenges. For example, the diversity and inconsistency of EHR data may affect the performance of the model, and data privacy and security issues are also important considerations. In addition, how to achieve efficient training and inference of RAG models in practical applications is also an urgent problem that needs to be solved.

In the study [1], the author used the Llama 213B model for zero sample prompting in the context of effectively extracting key clinical information from a large amount of data in EHR [1]. The study tested the ability of the Llama 2 model to summarize and extract risk factors for malnutrition and weight loss from nursing progress notes in residential elderly care, both individually and in combination with RAG.

In another study [2], the author proposes a novel framework called REALM, aimed at addressing some key issues in existing EHR data analysis models in clinical prediction tasks, and optimizing some existing capabilities in the EHR field, such as enhancing medical context understanding, integrating unstructured data, and improving clinical prediction performance.

In the study [3], the application of different Large Language Models (LLMs) in text to SQL tasks is reviewed, as well as how RAG technology enhances the performance of these models. The major models involved in the article mainly include OpenAI's GPT-3.5 Turbo, OpenAI's GPT-4 Turbo, Google's GeminiPro 1.0, Aerospace's Claude 2.1, Mistral AI's Mixture 8x7B, Mistral AI's Mixture Medium, etc., fully demonstrating the potential of RAG technology in practical problem-solving, especially in scenarios that require processing large amounts of complex data.

This article will provide a detailed overview of the latest research progress and application examples of the RAG model in the field of EHR, analyze its advantages and limitations, and explore future development directions and research hotspots. It can provide valuable references for relevant researchers and practitioners and promote the widespread application and development of RAG technology in the field of EHR.

2. Method

2.1 Introduction of RAG

Retrieval-Augmented Generation is a technical framework that combines Large-scale Language Models (LLMs) with external knowledge retrieval to enhance the accuracy of question answering and content generation. It improves the generation of more accurate and relevant responses by retrieving information from data sources and using this information as context input to the language model.

The RAG workflow typically includes three main steps, namely information retrieval, information integration and generative language modeling. First, the system retrieves relevant information from a large knowledge base or data source. Then, it integrates the retrieved information with the current generation task to form a richer context. Finally, based on this context, the generator uses the language model to produce the final answer or text.

The core of RAG technology lies in its ability to effectively utilize external knowledge sources to overcome the limitations of traditional language models in specific domain tasks. This includes reducing model hallucinations and improving the accuracy and relevance of generated content. Additionally, RAG can handle knowledge from specialized fields, providing more comprehensive natural language processing solutions.

2.2 The Application of RAG in Abstract Generation

In the experiment of Alkhalaf et al. [1], the authors used RAG technology to extract key clinical information from EHR. In the experiment, the author first divided the nursing notes into fixed size blocks (600 characters) and parsed structured data (e.g. demographics and weight scales). Then they used the sentence vector transformation model in the Huggingface library to encode the data into dense vectors. And the encoded vector was stored in the vector storage for subsequent fast similarity search. Then the Maximum Marginal Correlation Retrieval (MMR) algorithm was used to retrieve relevant documents from vector storage. Then the authors sent the retrieved documents to the Llama 2 model and generated a summary with specific prompt instructions. Finally, the BitsandBytes library was employed to quantify the weights of the Llama 2 model to adapt to GPU memory limitations. In this article, the RAG method significantly improved the accuracy of the model in summary generation tasks, from 93.25% to 99.25%.

In Walid Saba et al.'s study [2], the authors applied RAG technology to question answering mode to achieve EHR abstract generation. In the experiment, the author used a publicly available dataset of over 2 million clinical notes - MIMIC-III. Firstly, the author segmented the EHR data and established a vector database index. Then, through question loops, relevant text fragments were retrieved, and answers were generated using LLMs. Finally, they plan to evaluate the accuracy and completeness of generated abstracts through expert evaluation and calculation of indicators such as ROUGE and BLEU. However, the author did not provide the final effect of the model in the article, so the effectiveness of RAG in the field of EHR summary generation based on question answering mode needs further research.

2.3 The Application of RAG Model in the Field of Epidemiological Informatics

In the article by Ziletti [3], the authors discuss how to improve the efficiency of querying patient health status and medical utilization by combining text with RAG technology to process real-world health data.

In this article, the author first uses different LLM prompts to transform natural language problems into SQL queries. Secondly, the medical coding step is introduced to map medical entities to precise clinical oncology terms. Finally, using RAG technology, the dataset is used as an external knowledge base to improve SQL generation by extracting relevant problem SQL pairs. Finally, the DE SynPUF dataset was used to evaluate the performance of the model using accuracy (Acc) and executable performance (Exec) as evaluation metrics.

The results of this article show that RAG significantly improves the performance of LLMs in different problems, especially when using the most similar problem (RAG top1). This also strongly proves that RAG has a very broad development and application prospect in the field of EHR.

2.4 The Application of RAG Model in the Field of Multimodal EHRs Analysis

In Zhu Yinghao et al.'s paper [4], a detailed analysis was conducted on the application prospects of RAG technology in enhancing multimodal EHRs. Specifically, the article explores how to overcome the shortcomings of existing models in medical background knowledge through retrieval enhanced generation methods, and proposes an adaptive multimodal fusion network framework, REALM, to integrate extracted knowledge with multimodal EHR data. This framework utilizes clinical notes and multivariate time series EHR data and improves clinical prediction ability by combining Knowledge Graphs (KGs).

In addition, the article also evaluated the performance of REALM through experiments on the MIMIC-III dataset. The research results showed that the REALM framework achieved significant performance improvements in predicting hospitalization mortality and 30-day readmission on the MIMIC-III dataset compared to the baseline model. Specifically, significant improvements were observed in evaluation metrics such as Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision Recall Curve (AUPRC) , minimum value of precision and sensitivity (min (+P, Se)), and F1 score. Through the ablation study, the article demonstrates that RAG enhancement has significant performance improvements for both time series and text modalities. This indicates that combining external medical knowledge can effectively capture more complex medical background knowledge, thereby improving the accuracy of clinical predictions, demonstrating the superior performance of the REALM framework compared to baseline models.

3. Discussion

Overall, there has been noteworthy progress in the application of RAG in the field of EHRs. This progress is particularly evident in the areas of abstract generation and multimodal EHR analysis, where RAG has shown significant potential in reducing hallucinations and enhancing the timeliness of LLMs. By leveraging external knowledge sources, RAG effectively mitigates the common issue of hallucinations, thereby improving the accuracy and reliability of generated content. Additionally, the incorporation of multimodal data sources, such as clinical notes and time-series data, has allowed for more comprehensive and nuanced analyses, ultimately leading to better clinical predictions and decision-making support.

Despite these advancements, it is important to recognize that RAG technology is still relatively new. As a result, the current understanding of its full capabilities and limitations in this field is not yet complete. The research into RAG technology, while promising, remains in its nascent stages, and there is much to be explored and validated. The application and implementation of RAG in the EHR field are not yet mature, facing several challenges that need to be addressed. These challenges include the integration of diverse data sources, ensuring the quality and relevance of retrieved information, and optimizing the performance of LLMs within the RAG framework.

Firstly, it is not difficult to find from different literature that RAG often has qualified generalization ability when dealing with single problems in the field of EHR. However, in future practical applications, RAG models may need to simultaneously or sequentially handle multiple problems in different fields, which means that RAG models need to be able to handle EHR data from different sources and formats. However, there is currently not enough in-depth research on the ability of RAG to handle multiple formats of data. And the EMERGE framework proposed by Zhu et al. [5] also has dependencies on specific datasets. Therefore, future researchers can delve deeper into this field to improve the multimodal data processing ability of RAG models, their ability to combine with knowledge graphs, further enhancing the generalization ability of RAG models based on some advanced technologies such as domain adaptation [6, 7].

Secondly, for the specific field of medicine, the RAG model also has another flaw, which is the interpretability of the model. In the field of EHR, the decision-making process of the model needs to be transparent and interpretable, making it easy for doctors and patients to understand the predicted results of the model and make corresponding targeted treatments. However, the RAG model is not outstanding in this regard. Although Zhu et al. have attempted to improve the interpretability of the model by matching medical entities with external knowledge graphs [5], this direction will still be an urgent challenge for the

RAG model to be applied in the EHR field in the future. Therefore, it is important to consider incorporating advanced interpretability algorithms [8] with HER in future research.

Thirdly, in order for the RAG model to be successfully applied in the EHR field, further optimization is needed. In the article by Basu et al. [9], the recall rate of Onco Retriever is 78%, and any information omission in the medical field can have a significant impact. Therefore, the model needs further optimization. The RECTIFIER model proposed by Ozan Unlu et al. [10] also has problems such as losing patient clinical background information and ignoring key clinical details, and further optimization of the model is needed.

4. Conclusion

This paper provides a comprehensive review related to the application of RAG in HER. It can be found that RAG technology has demonstrated significant progress in the field of EHR, particularly in summary generation and multimodal EHR analysis. RAG technology has shown excellent performance in reducing hallucinations and improving the timeliness of LLMs. By combining external knowledge sources, RAG effectively reduces common hallucination issues and improves the accuracy and reliability of generated content. At the same time, by combining multimodal data sources such as clinical notes and time series data, the analysis becomes more comprehensive and detailed, thereby promoting better clinical prediction and decision support.

Although current research is promising, it is still in its early stages and requires further exploration and validation. The application and implementation of RAG in the field of EHR are not yet mature and face several challenges, including integrating diverse data sources, ensuring the quality and relevance of retrieval information, and optimizing LLM performance within the RAG framework. Future research needs to focus on addressing these issues. Meanwhile, researchers should continue to explore the combination of RAG and knowledge graph to further enhance the generalization ability of RAG models. Through continuous innovation and collaboration, RAG technology is expected to achieve wider applications in the medical field, ultimately improving patient treatment outcomes and healthcare service efficiency.

References

[1] Alkhalaf 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, 2024: 104662.

[2] Saba W, Wendelken S, Shanahan J. Question-Answering Based Summarization of Electronic Health Records using Retrieval Augmented Generation. arXiv preprint arXiv:2401.01469, 2024.

[3] Ziletti A, D'Ambrosi L. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. arXiv preprint arXiv:2403.09226, 2024.

[4] Zhu Y, et al. REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. arXiv preprint arXiv:2402.07016, 2024.

[5] Zhu Y, et al. EMERGE: Integrating RAG for Improved Multimodal EHR Predictive Modeling. arXiv preprint arXiv:2406.00036, 2024.

[6] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. Advances in Neural Information Processing Systems, 2006, 19.

[7] Qiu Y, Hui Y, Zhao P, Wang M, Guo S, Dai B, Dou J, Bhattacharya S, Yu J. The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process. Fuel, 2024, 372: 132162.

[8] Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, Srivastava M, et al. Interpretability of deep learning models: A survey of results. In 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), 2017, pp. 1-6. IEEE.

[9] Gupta SK, et al. Onco-Retriever: Generative Classifier for Retrieval of EHR Records in Oncology. arXiv preprint arXiv:2404.06680, 2024.

[10] Unlu O, et al. Retrieval-Augmented Generation–Enabled GPT-4 for Clinical Trial Screening. NEJM AI, 2024: AIoa2400181.