Retrieval-Augmented Generation with Large Language Models in Nephrology

Subjects: Medical Informatics

Contributor: Jing Miao , Charat Thongprayoon , Supawadee Suppadungsuk , Oscar A. Garcia Valencia , Wisit Cheungpasitporn

The integration of large language models (LLMs) into healthcare, particularly in nephrology, represents a significant advancement in applying advanced technology to patient care, medical research, and education. These advanced models have progressed from simple text processors to tools capable of deep language understanding, offering innovative ways to handle health-related data, thus improving medical practice efficiency and effectiveness. A significant challenge in medical applications of LLMs is their imperfect accuracy and/or tendency to produce hallucinations—outputs that are factually incorrect or irrelevant. This issue is particularly critical in healthcare, where precision is essential, as inaccuracies can undermine the reliability of these models in crucial decision-making processes. To overcome these challenges, various strategies have been developed. One such strategy is prompt engineering, like the chain-of-thought approach, which directs LLMs towards more accurate responses by breaking down the problem into intermediate steps or reasoning sequences. Another one is the retrieval-augmented generation (RAG) strategy, which helps address hallucinations by integrating external data, enhancing output accuracy and relevance. Hence, RAG is favored for tasks requiring up-to-date, comprehensive information, such as in clinical decision making or educational applications.

large language models (LLMs) nephrology chronic kidney disease artificial intelligence retrieval-augmented generation (RAG)

1. What Is the RAG System?

The RAG approach is a method used in natural language processing and machine learning that combines the strengths of retrieval-based and generative models to improve the quality of generated text [1][2]. This approach is particularly useful in tasks such as question answering, document summarization, and conversational agents. In the dynamic field of medicine, the unique capability of the RAG system to access external medical databases in real time allows the LLM to base its responses on the latest research, clinical guidelines, and drug information [3][4].

To generate more accurate and contextually relevant responses, the RAG approach combines the strengths of two components including the retrieval and generation components. The former component is responsible for fetching relevant information or documents from a large database or knowledge source provided to the LLMs. The retrieval is typically based on the input query or context, aiming to find content that is most likely to contain the information needed to generate an accurate response. The latter component takes the input prompt along with the retrieved documents or information from the retrieval component and generates a response. The generation component uses the context provided by the retrieved documents to inform its responses, making them more accurate, informative, and contextually relevant.

The RAG approach is particularly beneficial in scenarios where the model needs to provide information that may not have been present in its training set or when the information is continually updated. By grounding the responses in factual data, the RAG approach effectively reduces the occurrence of inaccuracy or hallucinations. However, the success of RAG depends on the quality and timeliness of the external data sources, and integrating these sources introduces additional technical complexities. Complementing these approaches is the process of fine-tuning, which involves adapting a pre-trained model to specific tasks or domains. This enhances the model's capacity to process certain types of queries or content, thereby improving its efficiency and specificity for certain domains. While this method improves the model's performance in specific areas, it also poses the risk of over-fitting in certain datasets, potentially limiting its broader applicability and increasing the demands on training resources.

2. Current Research Regarding the Application of RAG in Medical Domain

A recent study experimentally developed a liver disease-focused LLM model named LiVersa, incorporating the RAG approach with 30 guidelines from the American Association for the Study of Liver Diseases. This integration was intended to enhance LiVersa's functionality. In the study, LiVersa accurately answered all 10 questions related to hepatitis B virus treatment and hepatocellular carcinoma surveillance. However, the explanations provided for three of these cases were not entirely accurate [5]. Another study introduced Almanac, an LLM framework enhanced with RAG functions, which was specifically integrated

with medical guidelines and treatment recommendations [6]. This framework's effectiveness was evaluated using a new dataset comprising 130 clinical scenarios. In terms of accuracy, Almanac outperformed ChatGPT by an average of 18% across various medical specialties. The most notable improvement was seen in cardiology, where Almanac achieved 91% accuracy compared to ChatGPT's 69% [6]. Moreover, they evaluated the performance of Almanac against conventional LLMs (ChatGPT-4 [May 24, 2023 version], BingChat [June 28, 2023], and Bard AI [June 28, 2023]) by testing the LLMs with a new dataset comprising 314 clinical questions across nine medical specialties. Almanac demonstrated notable enhancements in accuracy, comprehensiveness, user satisfaction, and resilience to adversarial inputs when compared to the standard LLMs [7]. A recent investigation introduced a RAG system named RECTIFIER (RAG-Enabled Clinical Trial Infrastructure for Inclusion Exclusion Review), assessing its efficacy against that of expert clinicians in a clinical trial screening [8]. The comparison revealed a high concordance between the responses from RECTIFIER and those from expert clinicians, with RECTIFIER's accuracy spanning from 98% to 100% and the study staff's accuracy from 92% to 100%. Notably, RECTIFIER outperformed the study staff in identifying the inclusion criterion of "symptomatic heart failure", achieving an accuracy of 98% compared to 92%. In terms of eligibility determination, RECTIFIER exhibited a sensitivity of 92% and a specificity of 94%, whereas the study staff recorded a sensitivity of 90% and a specificity of 84%. These findings indicate that integrating a RAG system into GPT-4-based solutions could significantly enhance the efficiency and cost effectiveness of clinical trial screenings [8].

The RAG's strengths lie in its access to current information and its ability to tailor relevance. By utilizing the most recent data, the likelihood of offering outdated or incorrect information is greatly reduced. However, this approach also presents several challenges. The effectiveness of RAG's responses is heavily dependent on the quality and currency of the data sources it uses. Adding RAG to LLMs also introduces an extra layer of complexity, which can complicate implementation and ongoing management. Moreover, there is a risk of retrieval errors. Should the retrieval system malfunction or fetch incorrect information, it could result in inaccuracies in the output it generates.

3. The Potential Applications of RAG in Nephrology

The RAG integration is also valuable in nephrology, where staying abreast of the latest developments is crucial. This integration of current, validated data from external sources significantly reduces the likelihood of the LLMs providing outdated or incorrect information.

3.1. Integrating Latest Research and Guidelines

The RAG approach has the unique capability to dynamically integrate the most recent findings from nephrology-related sources into the model's outputs. This includes new research from nephrology journals, results from the latest clinical trials, or any updates in treatment guidelines. By doing so, the RAG approach ensures that LLMs are not only up-to-date but also highly relevant and accurate in the field of nephrology. For instance, consider a scenario where a nephrology specialist or an internist is seeking information about the latest management strategies for polycystic kidney disease (PKD). In such cases, the RAG can actively search for, retrieve, and incorporate information from the most recent guidelines and treatment protocols, such as the KDIGO 2023 clinical practice guideline for autosomal dominant polycystic kidney disease (ADPKD), and studies published in the PubMed database. This process involves not just accessing this information but also synthesizing it in a way that is coherent and directly applicable to the query at hand.

By utilizing RAG, the physician is thus provided with information that is not only current but is also directly relevant to their specific inquiry. This approach is especially valuable in a field like nephrology, where advancements in research and changes in treatment protocols can have a significant impact on patient care. The ability of RAG to provide the latest knowledge helps healthcare professionals stay informed and make well-founded decisions in their practice.

3.2. Case-Based Learning and Discussion

Employing RAG in educational settings can significantly enhance the learning process by incorporating detailed and real-life case studies into lectures, discussions, or interactive learning modules. This application of RAG is particularly useful in complex and dynamic fields like medicine. Take, for example, the education of medical students on the topic of complex electrolyte imbalances in chronic kidney disease (CKD). The RAG approach can be utilized to access and reference specific, real-world case reports or clinical scenarios relevant to this topic. By doing so, it can provide students with practical, tangible examples that illustrate the theoretical concepts they are learning. This not only aids in a deeper understanding of the subject matter but also helps students appreciate the real-world implications and applications of their knowledge.

Moreover, RAG's ability to retrieve the latest studies and reports ensures that the educational content is not only rich in practical examples but also current. This is especially vital in medical education, where staying abreast of the latest research and clinical practices is crucial. By integrating up-to-date case studies and scenarios, RAG can help create a more engaging

and informative educational experience, preparing students for the challenges they will face in their medical careers. This approach can be extended to other complex medical topics, making learning more interactive, relevant, and evidence-based.

3.3. Multidisciplinary Approach

In situations where a multidisciplinary perspective is essential, RAG proves to be particularly valuable as it can draw upon a wide array of medical disciplines to offer a more comprehensive understanding. This capability is critical in treating conditions that intersect multiple areas of healthcare. Consider the case of a patient suffering from diabetic nephropathy, for instance. This condition, being at the crossroads of diabetes and kidney health, requires a nuanced understanding from several medical specialties. The RAG system can effectively consolidate relevant information from endocrinology, focusing on diabetes management strategies; from cardiology, addressing the cardiovascular risks associated with the condition; and from nephrology, providing insights into preserving renal function.

By integrating this diverse information, the RAG system can greatly assist healthcare professionals in developing a holistic and multifaceted treatment plan. This approach ensures that all aspects of the patient's condition are considered, leading to more effective and comprehensive patient care. Such an integrated approach is beneficial not just in diabetic nephropathy but in any complex medical condition where multiple body systems are affected or where various specialties need to collaborate for optimal patient management. The ability of RAG to seamlessly merge insights from different medical fields into a cohesive whole enhances its utility in planning and implementing effective treatment strategies.

4. Creation of a CKD-Specific Knowledge Base for RAG

To illustrate the process of creating a customized ChatGPT model with a RAG strategy, we will use the field of nephrology as a reference, specifically focusing on CKD due to its prevalence in nephrology encounters (**Figure 1**). This example will serve to demonstrate the steps and considerations involved in tailoring a ChatGPT model to a specific medical specialty, incorporating a specialized knowledge base. The aim is to enhance the model's responses with precise, specialized knowledge, in this case, centered around CKD, guided by insights from the KDIGO 2023 Clinical Practice Guideline [9]. Below is a detailed breakdown of the steps involved in this process.

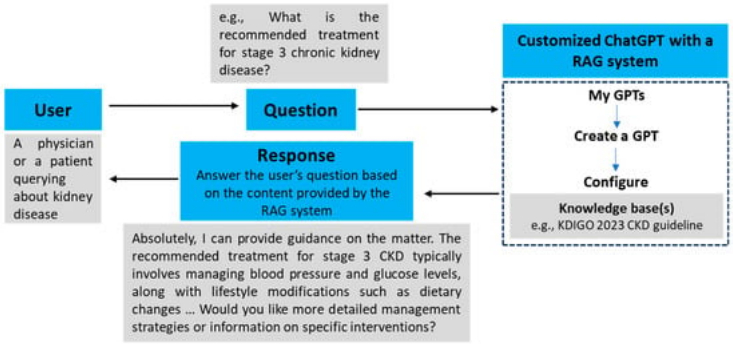

Figure 1. The process of creating a customized ChatGPT model with the retrieval-augmented generation (RAG) strategy in nephrology.

4.1. Creation of a CKD-Focused Retrieval System

This process involves the careful selection of knowledge sources, integration of guidelines, and regular updates to ensure accuracy and relevancy. The first step is to meticulously select a comprehensive database rich in information about CKD. This database should draw from a range of reliable sources, such as peer-reviewed academic journals, reports from clinical trials, and authoritative nephrology textbooks. A key focus is placed on incorporating the KDIGO 2023 CKD guidelines [9], which are recognized for their currency and authority in the field.

Next, it is vital to directly integrate these KDIGO 2023 guidelines into the chosen database by creating a customized ChatGPT model (**Figure 2**). This process involves navigating to "My GPTs" and selecting "Create a GPT". Following this, we have the opportunity to customize/configure our GPT by entering a name, description, and instructions, and by uploading the knowledge bases(s) we wish to embed within the model. We can choose to restrict access to the model by selecting one of the following options: "Only me", "Anyone with a link", or "Everyone". Once customized, the GPT will be accessible under "My GPTs", where it will produce responses utilizing the incorporated database(s).

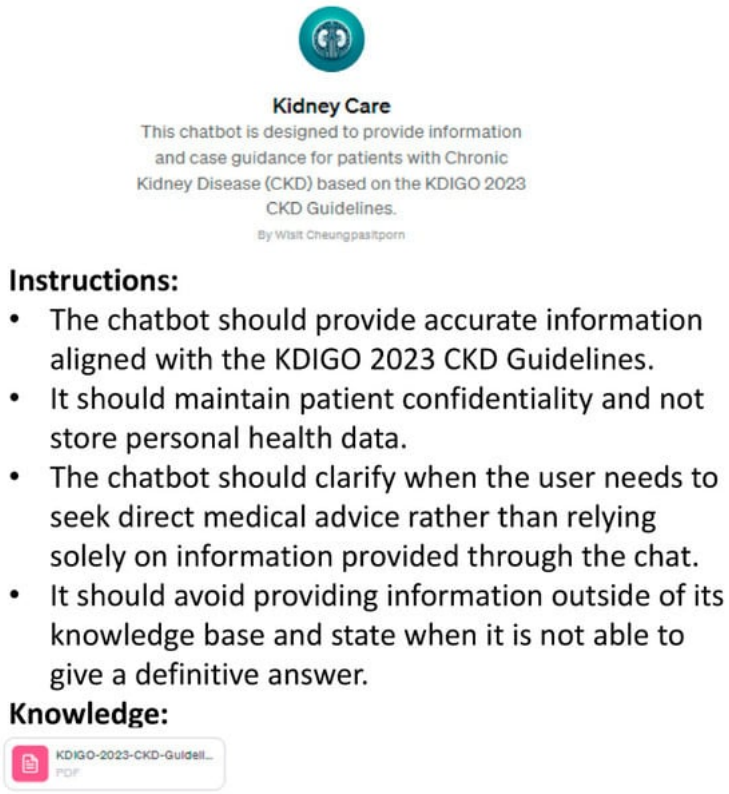

Figure 2. The creation of a CKD-specific knowledge base by customizing GPT-4 with the retrieval-augmented generation (RAG) approach.

This integration covers the detailed aspects of CKD, including diagnosis, staging, management, and treatment protocols. Such incorporation ensures that the model's responses are in line with the most recent and accepted clinical practices. While ChatGPT operates based on its internal knowledge gained during training, RAG takes this a step further by dynamically incorporating external information into the generation process. The integration of a retrieval component in RAG could theoretically enhance ChatGPT by providing it access to a wider range of current information and specific data not covered during its training.

4.2. Development of a CKD-Focused Retrieval System

The RAG system, specialized for CKD, is specifically configured to identify and respond to CKD-related queries accurately. It is adept at grasping the intricacies of CKD, including its various stages, the comorbid conditions often accompanying it, and the diverse methods of treatment available. Additionally, the system is fine-tuned for both speed and relevance, ensuring rapid and efficient access to relevant information from the comprehensive CKD database when processing queries. This optimization guarantees prompt and pertinent responses tailored to the specifics of CKD. Moreover, establishing a system for continuous updates to the database is crucial. This involves regularly reviewing and including new research findings, updated medical guidelines, and emerging treatment methods in nephrology. Keeping the database up to date guarantees that the information remains both current and authoritative, making it a reliable foundation for the model's knowledge base.

4.3. Integration with the Customized GPT-4 Model

Integrating the customized GPT-4 model with the CKD retrieval system involves establishing strong and secure API (Application Programming Interface) connections. Firstly, it focuses on creating a robust connection that allows for the seamless flow of data between the customized ChatGPT model and the CKD retrieval system. This connection must be secure to protect sensitive medical information and ensure data integrity. Secondly, the customized ChatGPT model undergoes fine-tuning to harmonize the in-depth CKD information with its innate natural language processing abilities. This fine-tuning is critical to ensure that the model not only provides responses that are accurate and rich in CKD-specific information but also maintains clarity and appropriateness in the context of the user's query.

Through this integration, the model becomes capable of delivering responses that are not just factually correct but also tailored to the specific context of the query, whether it is a patient's inquiry, a healthcare professional's detailed question, or an

educational scenario. This ensures that the model's outputs are highly relevant, understandable, and useful for various users, ranging from medical practitioners and students to patients seeking information about CKD.

4.4. Customized Response for CKD Inquiries

The integration of a customized GPT-4 model with a CKD-specialized RAG system brings a significant advancement in handling CKD-related inquiries. This integration leverages sophisticated algorithms to ensure that the ChatGPT model precisely recognizes the context and specific details of queries related to CKD, leading to highly relevant and tailored responses. This process operates on multiple levels, including contextual understanding, relevance of responses, access to updated information, and dynamic information integration.

Through this integrated approach, the ChatGPT model becomes a powerful tool for providing accurate, up-to-date, and highly specific responses to a wide range of CKD-related inquiries. This capability is particularly valuable for healthcare professionals seeking quick and reliable information, patients looking for understandable explanations of their condition, and researchers needing the latest data in the field of nephrology.

4.5. Rigorous Testing with CKD Scenarios

The system undergoes comprehensive testing in a variety of CKD situations. This testing encompasses a spectrum of patient histories, various stages of CKD, and the intricacies involved in treatment plans. Such extensive testing is crucial for confirming the model's reproductivity and its ability to adapt to diverse clinical conditions. The feedback obtained from these rigorous tests is instrumental to the ongoing enhancement of the system. It aids in refining the precision of information retrieval and boosting the effectiveness of how the ChatGPT model works in conjunction with the CKD database. This process of continuous improvement ensures the system remains reliable and effective in addressing the complex needs of CKD management.

4.6. Regular System Monitoring and Updating

The system's performance in providing accurate and relevant CKD information is consistently monitored. This includes assessing the accuracy of responses, the relevance of information provided, and the speed of retrieval. Moreover, the CKD database is regularly updated with the latest research, guidelines, and treatment protocols, ensuring the model's responses remain current and authoritative.

References

1. Merritt, R. What Is Retrieval-Augmented Generation, Aka RAG? Available online: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/#:~:text=Generation%20(RAG)%3F-,Retrieval%2Daugmented%20generation%20(RAG)%20is%20a%20technique%20for%2 (accessed on 15 November 2023).
2. Guo, Y.; Qiu, W.; Leroy, G.; Wang, S.; Cohen, T. Retrieval augmentation of large language models for lay language generation. J. Biomed. Inform. 2023, 149, 104580.
3. Luu, R.K.; Buehler, M.J. BioinspiredLLM: Conversational Large Language Model for the Mechanics of Biological and Bio-Inspired Materials. Adv. Sci. 2023, e2306724.
4. Wang, C.; Ong, J.; Wang, C.; Ong, H.; Cheng, R.; Ong, D. Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation. Ann. Biomed. Eng. 2023.
5. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.C.; Pletcher, M.J.; Lai, K. Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation. medRxiv 2023.
6. Zakka, C.; Chaurasia, A.; Shad, R.; Dalal, A.R.; Kim, J.L.; Moor, M.; Alexander, K.; Ashley, E.; Boyd, J.; Boyd, K.; et al. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. Res. Sq. 2023.
7. Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A.R.; Kim, J.L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.; et al. Almanac —Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI 2024, 1, AIoa2300068.
8. Unlu, O.; Shin, J.; Mailly, C.J.; Oates, M.F.; Tucci, M.R.; Varugheese, M.; Wagholikar, K.; Wang, F.; Scirica, B.M.; Blood, A.J.; et al. Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. medRxiv 2024.

9. KDIGO 2023 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Available online: https://kdigo.org/guidelines/ckd-evaluation-and-management/ (accessed on 1 July 2023).