Applying Retrieval-Augmented Generation on Open LLMs for a Medical Chatbot Supporting Hypertensive Patients

Gianluca Aguzzi[1,†], Matteo Magnini[1,†], Giuseppe Pio Salcuni[1], Stefano Ferretti[1] and Sara Montagna[2]

[1]Department of Computer Science and Engineering, University of Bologna, Via dell'Università 50, Cesena, Italy
[2]Department of Pure and Applied Sciences, University of Urbino, Piazza della Repubblica 13, Urbino, Italy

Abstract

Disease management, especially for chronic conditions or the elderly, involves continuous monitoring, lifestyle adjustments, and frequent healthcare interactions, necessitating effective home-care ICT solutions. To address these needs, chatbot technology has emerged as a promising tool for supporting patients in managing their health autonomously. In this context, chatbots must provide timely and accurate information and continuous empathetic support to maintain patient engagement. Additionally, data privacy concerns necessitate avoiding third-party Natural Language Processing and Generation services.

To meet these needs, in this paper we propose the development of a chatbot to support patients in managing chronic conditions, focusing on hypertension. Particularly, we utilise open-source large language models to avoid proprietary systems due to privacy requirements. Given that their performance, based on state-of-the-art metrics, do not compete third-party services, we incorporate retrieval augmented generation (RAG) techniques, building a knowledge base with input from medical professionals to enhance model performance. We evaluated seven open-source models, including two specifically trained in the medical domain. Our results indicate that RAG significantly improves performance, surpassing that of specialised medical-domain models without RAG. This approach offers a promising solution for managing chronic conditions independently and securely.

Keywords

Chronic Disease Self-management, Large Language Models, Retrieval-Augmented Generation

1. Introduction

Chronic disease management presents a substantial challenge for both healthcare systems and patients. Conditions like hypertension require continuous monitoring, lifestyle adjustments, and often involve significant healthcare costs. This burden is amplified by the need for frequent interaction with healthcare professionals, leading to increased wait times and potential access barriers for patients. To address this, we propose the development of a chatbot designed to support patients in the self-management of chronic conditions, with a focus on hypertension. The goal is to empower hypertensive patients to manage their condition more independently by providing them with timely, accurate, and empathetic guidance, particularly aimed at periodically acquiring patient vital signs and at maintaining a healthy lifestyle [1, 2].

Two critical requirements emerge. First, the interaction between the patient and the chatbot must be as empathetic as possible to ensure patients remain motivated and engaged in managing their condition. At the same time, the information provided by the chatbot must be highly accurate, as there is no healthcare professional directly mediating the conversation. Second, given that patients are likely to share personal health data during these interactions, the chatbot must comply with data privacy

HC@AIxIA 2024: 3rd AIxIA Workshop on Artificial Intelligence For Healthcare

[†]These authors contributed equally.

gianluca.aguzzi@unibo.it (G. Aguzzi); matteo.magnini@unibo.it (M. Magnini); giuseppepio.salcuni@studio.unibo.it (G. P. Salcuni); s.ferretti@unibo.it (S. Ferretti); sara.montagna@uniurb.it (S. Montagna)
0000-0002-1553-4561 (G. Aguzzi); 0000-0001-9990-420X (M. Magnini); 0009-0001-8904-0695 (G. P. Salcuni); 0000-0002-1911-4708 (S. Ferretti); 0000-0001-5390-4319 (S. Montagna)

regulations, which precludes the use of third-party systems for natural language processing (NLP) and natural language generation (NLG).

Considering these requirements, we look to large language models (LLMs) as the core technology for our chatbot. Their demonstrated ability to produce trustable, reliable and emphatic text makes them an ideal choice for this application. For instance, [3] shows that a set of patients, that receive responses in a public online forum both from physicians and from a LLM-based chatbot, preferred chatbot replies, rating both the quality and the empathy of the chatbot-generated responses higher than those of physicians. We focus on open LLMs, including both models pre-trained specifically for medical domains and general-purpose models, to avoid reliance on proprietary third-party services. To enhance the performance of these open LLMs, we explored the integration of retrieval augmented generation (RAG) techniques [4].

This approach involved constructing a knowledge base by collecting data from medical professionals and subsequently enriching this dataset using RAG techniques. This enabled us to perform a comprehensive comparison of various retrieval strategies and LLM models (both specialised and general-purpose). Our findings demonstrate that RAG is a vital technique for significantly improving the performance of these models, even surpassing the performance of specialised models in mainly all the tested cases.

The remainder of this paper is organised as follows. Section 2 provides background information and motivation for our work in the context of chronic disease management and the development of a chatbot for hypertensive patients; Section 3 describes the methods used in our study, including the RAG technique, and details the dataset used for our experiments; Section 4 presents the results of our study and discusses the implications of our findings; and Section 5 concludes the paper with a summary of our work and suggestions for future research.

2. Background and Motivation

LLM applications in the healthcare field span various contexts and serve multiple purposes. Over the past year, the number of research studies focusing on LLMs has grown exponentially, reflecting the increasing interest within the scientific community working on artificial intelligence (AI) applied to medicine. The adoption of LLMs has become a focal point for numerous investigations in this area.

Potential application areas, as outlined by [5], can be divided into three main categories:

Patient Care There is a general agreement that LLM-based chatbots may become a methodological tool assisting physicians or nurses, during their clinical practice, in various areas of medicine. As an example, they may support clinical decisions, by abstracting key results from literature. Or they can detect medical errors, by identifying discrepancies between diagnosis and treatment. On the patient side, they may be a crucial component for bootstrapping patient empowerment by providing trustworthy and emphatic answers to user queries. In this context, they must resemble a dialogue between the physician and the patient which is a key element to provide an effective and compassionate care. Moreover, they should be able to proactively suggest actions, reasoning on tracked patient activities and vital signs dynamic.

Research In research, an LLM-based chatbot may assist basic research by automating certain tasks, such as data analysis, acquisition and interpretation, summarising information, paraphrasing text, scientific literature search for medical knowledge and related work extraction.

Education In medical education, an LLM-based chatbot may be used to provide teaching material and as a tool for students who can benefit from interacting tutoring. In this context, noteworthy are the very good performances demonstrated in passing medical examinations.

In this study, we focus on the use of LLMs as an integral part of a chatbot designed for supporting hypertensive patients. This chatbot is designed to collect hypertension parameters, motivate patients with periodic messages suggesting healthy lifestyle changes, and assist them with any concerns related to their chronic condition [1, 2].

Three key requirements emerge for such an application:

1. Ethical concerns, including risks of privacy and security [6] must be addressed: third-party technology, such as ChatGPT, carries an inherent risk of compromising patient privacy, if patients enter test results, photos of their face, communication information, and more. All of this vital health information is collected and stored, potentially compromising patient privacy.
2. The system must be highly reliable, with no hallucinations or erroneous information: Before deploying LLMs in real-world medical environments, it is essential to ensure that models designed for healthcare are accurate, unbiased, and safe for patient use [7].
3. The system should communicate empathetically, motivating the patient, making them feel heard, and providing ongoing real-time support [3].

Many of today's top-performing LLMs are proprietary models with hundreds of billions of parameters trained on vast amounts of data. However, due to the first requirement, the immediate choice falls on open-source LLMs that can be deployed locally. While these models show promise in matching the performance of proprietary counterparts, they have notable limitations, such as generating false or irrelevant information and mistyped or incomplete sentences, which can undermine the trustworthiness, accuracy, and regulatory compliance of LLM-generated content. On the positive side, open-source models allow developers to access model weights directly, enabling hosting on their own infrastructure. Accordingly, to address the second and third requirements, the literature suggests two primary techniques: querying local databases to complete specific tasks through RAG and fine-tuning, for which technical details will be provided in Section 3. These approaches are recommended to improve model performance and enhance the conversational experience according to a domain-specific dataset.

2.1. Related Work

On the RAG in Medical Domain The RAG model represents a significant innovation in the integration of information retrieval and generative models, allowing access to a specific medical knowledge base for generating precise and contextually relevant responses, thus enabling a safer and more effective application and deployment of LLMs in healthcare. This technology is particularly useful in healthcare, a field that demands high precision and sensitivity where accuracy and specificity of information are not just metrics but directly affect patient care quality.

Most of the studies found in literature reports efforts in supporting the clinical decision-making in specific medical domain, by deploying clinical decision support systems based on LLMs that exploit RAG based on relevant national guidelines in diverse context. For example, in [8], a RAG approach is implemented to improve the accuracy of LLMs, ensuring their outputs are consistent with expert knowledge in the field of digestive diseases. [9] introduces a new LLM framework that combines clinical guidelines with RAG to enhance text interpretation for managing Hepatitis C Virus infection. The findings indicate that this integrated framework outperforms the baseline LLM GPT-4 Turbo model in delivering precise, guideline-specific recommendations.

However, implementing RAG in healthcare presents numerous challenges due to the inherent diversity of clinical practices across healthcare institutions. These variations are influenced by multiple factors, such as patient demographics, available resources, geographical context, and specific cultural sensitivities. Therefore, tailoring RAG solutions to local needs is crucial. The models must adapt to variables like resource availability, specific medical protocols, and ethical and cultural practices. In this context, RAG's flexibility becomes essential, requiring a modular and highly configurable architecture.

Fine-tuning in Medical Domain LLMs fine-tuning has demonstrated impressive results against medical benchmarks. The literature is rich with examples that demonstrate the successes of fine-tuning in various specialised medical domains, highlighting its effectiveness and versatility in enhancing model performance [10]. For instance, Wang et al. [11] deploy the Llama 2 base model to generate training sentences that incorporate clinical concepts drawn from standardised vocabularies, particularly focusing on rare diseases. This approach utilises resources like the Human Phenotype Ontology

(HPO) to ensure accurate concept normalisation. By aligning model outputs with established medical terminologies, the study aimed to mitigate issues related to underdiagnosis, misdiagnosis, and mistreatment.

In this paper we focus on experimenting the RAG since the current dataset size does not allow for efficient fine-tuning. Moreover, the RAG approach allow to update the knowledge base dynamically, ensuring that the model remains relevant with recent information.

3. Materials & Methods

3.1. Fine-tuning

LLMs are initially trained on extensive, general-purpose text corpora with the aim of predicting the next token in a sequence, adjusting model parameters to maximise the likelihood of accurate predictions. This process, known as *pre-training*, equips the model with a broad understanding of language and serves as a foundation for fine-tuning.

During fine-tuning, the model is further trained on a smaller, domain-specific dataset, allowing it to adapt to the nuances of a particular task or field. This approach is computationally efficient compared to training from scratch, as it leverages the model's pre-existing knowledge while refining it for specific applications. Fine-tuning adjusts the model's weights based on new data to better align the model with the intended task such as summarisation, translation, or domain-specific text generation.

Full Fine-Tuning This method updates all the model's parameters to specialise its performance for a specific task. While this enhances model capabilities, it is resource-intensive, requiring significant computational power and memory, comparable to initial training.

Parameter-Efficient Fine-Tuning (PEFT) A more efficient approach, PEFT [12, 13][1] updates only a subset of the model's parameters, freezing the rest. This reduces memory requirements while retaining the model's general linguistic knowledge. Techniques like low-rank adaptation (LoRA) [14] refine smaller weight matrices to minimise resource usage, and Quantized LoRA (QLoRA) [15] further compresses memory demands by lowering the precision of the adapter weights.

By selecting the appropriate fine-tuning method, models can be optimised for specific use cases, maintaining performance while minimising resource consumption.

3.2. RAG

Traditionally, LLMs generate responses solely based on patterns and information learned during the training phase. However, these models are inherently limited by the data on which they were trained, often leading to responses that may lack depth or specific knowledge. RAG overcomes this limitation by drawing on external data during the response generation process.

The functioning of RAG involves two phases: first retrieving relevant information from a large dataset or knowledge base in response to a query, second using that information to inform and guide the generation of the response. This approach enables software agents – such as chatbots – to provide more accurate and context-specific answers supplementing the model's internal knowledge with relevant external information, such as private documentation, PDF files, or SQL databases. Figure 1 summarises the key components of a RAG system.

The retriever in a RAG system identifies relevant information to help answer a query. It begins by loading documents, splitting them into smaller fragments (a.k.a., chunks), and converting these chunks into embedding vectors using specialised algorithms. These vectors are stored in an indexed knowledge base for efficient retrieval.

[1] https://github.com/huggingface/peft

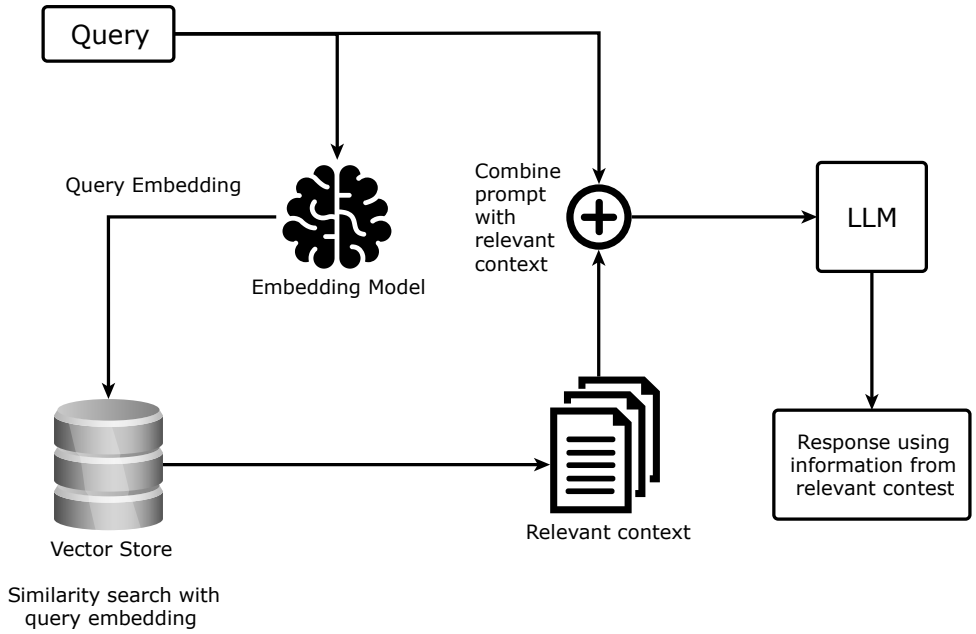


Figure 1: Principal components and phases of a RAG system.

When a new query is processed, the system generates a query vector and matches it with stored document vectors using vector similarity techniques. Two main types of embeddings are used:

Sparse Embeddings These rely on keyword matches, using algorithms like TF-IDF [16] and BM25 [17]. They are computationally efficient but may struggle with synonyms and semantic meaning.

Dense Embeddings These are generated by language models like BERT [18], capturing deeper semantic relationships and enabling retrieval based on meaning rather than exact words.

Hybrid approaches, combining both methods, can optimise retrieval by balancing speed and semantic depth, leading to more accurate and context-aware responses.

The generator is a language model that produces the final text in response to a query. It does not work in isolation but relies on the context provided by the retriever to guide its response, ensuring that the output is both plausible and detailed. Once the most relevant passages are retrieved, the generator synthesises this information and expresses it in natural language to produce the final answer.

3.3. RAG vs. Fine-tuning

RAG and fine-tuning represent two distinct approaches to enhancing foundational models, and their combination can yield significant performance improvements.

Knowledge Integration vs. Task Specialisation RAG integrates external knowledge into the generation process, providing the model with greater versatility and the ability to stay current. In contrast, fine-tuning hones the model for a specific task, thereby improving task-specific accuracy and efficiency.

Dynamic vs. Static Learning RAG facilitates dynamic access to up-to-date external data, ensuring the model remains relevant with recent information. Fine-tuning, however, is a static method, updating the model solely based on the most recent training cycle.

Generalisation vs. Customisation RAG preserves the model's generality by enriching it with external data, thus increasing adaptability across multiple tasks. Fine-tuning, in contrast, tailors the model for a specific use case, which may limit its performance in more generalised applications.

Resource Demands RAG requires a continuous retrieval and integration mechanism for external data, potentially making it resource-intensive during runtime. Fine-tuning, while resource-heavy during the training phase, does not demand additional resources during deployment.

3.4. Chosen Approach

In our study, we opted to experiment with RAG techniques to enhance the performance of open LLMs for the development of a chatbot supporting hypertensive patients. We selected different retrieval strategies and many LLMs, including both general-purpose and medical-domain-specific models, to evaluate the impact of RAG on model performance. We chose RAG over fine-tuning for several reasons. First, we plan to improve our previous work on the hypertension chatbot [1]. Because the users of the chatbot are intended to be numerous and diverse, we need a system that can adapt to different contexts and provide accurate information. With this respect, RAG is a more suitable approach, as it dynamically allows the model to access a wide range of external data sources. Second, fine-tuning requires more extensive computational resources, which we currently lack. Third, we want to avoid unpleasant behaviours such as model drift (e.g., due to outdated knowledge), hallucinations, and biases that can occur during fine-tuning. For all these reasons, we believe that RAG is the most appropriate approach for our study.

Concerning the retrievers, we tested three main strategies:

Base Retriever This strategy retrieves information based on vector similarity – maximum marginal relevance [19] – with the query. To generate the embeddings, we used the nomic-embed-text model [20] from Ollama available on [21].

Multi-Query Retriever To overcome the limits of the base retrieve, this method uses multiple queries to retrieve information, enhancing the diversity and relevance of the retrieved data. The queries are generated by an LLM – Llama3.1 [22] – based on the original query, then for each query, the retriever retrieves relevant chunks.

Ensemble Retriever This strategy combines the outputs of multiple retrievers to enhance retrieval performance. We used two retrievers – the base retriever and a second retriever based on BM25 – to generate the final response. Finally, the results are sorted using the reciprocal rank fusion algorithm [23] to select the most relevant information.

The documents that we use for the RAG consist of a dataset of a collection of question-answer pairs extracted from medical consultations. The dataset has 1,473 question-answer pairs, with each pair consisting of a question that a patient might pose to a chatbot and the corresponding answer that the chatbot should generate. The questions cover a wide range of topics related to hypertension, including symptoms, causes, treatments, and lifestyle recommendations. All the questions and answers are in Italian, as the chatbot is intended for use in Italy. All the answers have been reviewed by medical professionals to ensure their accuracy and relevance to the questions.

The LLMs we evaluated consist both of general-purpose models (e.g., LLama 3.1, Qwen2, Mistral Nemo, Phi3, Gemma2) and medical-domain-specific models (e.g., Llama3.1-Medical, Qwen2-Medical)—see Table 1 for more details.

4. Evaluation

4.1. Experimental setup

We assessed the effectiveness of various RAG techniques using the RAGAS [2] framework. This framework provides a comprehensive suite for evaluating different metrics and employs a test set derived from our training set. Adhering to the RAGAS methodology, instead of partitioning the dataset into training and

[2] https://docs.ragas.io/en/stable/

Table 1
Evaluated LLMs for our study.

Model	Description
Llama3.1	A general-purpose LLM based on the Llama3 architecture with 8B parameters, trained on a diverse range of text data [24].
Llama3.1-Medical	A medical-domain-specific version of Llama3.1, fine-tuned on medical data to enhance its performance in healthcare applications: https://ollama.com/qordmlwls/llama3.1-medical.
Qwen2	A general-purpose LLM with 7B parameters, trained on 29 different languages to improve cross-lingual performance [25].
Qwen2-Medical	A medical-domain-specific version of Qwen2, fine-tuned on medical data to improve its performance in healthcare tasks: https://ollama.com/echelonify/med-qwen2.
Mistral-Nemo	A 12B parameter model with a large context window (128K tokens) developed by Nvidia: https://ollama.com/library/mistral-nemo.
Phi3	A relatively small LLM with 3B parameters, trained by Microsoft on filtered high-quality data [26].
Gemma2	A 9B parameter model based on Deepmind Gemini developed by Google [27].

test sets, we leveraged an external LLM (GPT-4o) to generate a test set of 20 question-context-answer triplets. These generated triplets maintain statistical relevance to the original dataset. We evaluated a range of state-of-the-art open-source LLMs, both specialised (medical domain) and general-purpose (see Section 3.4), with and without RAG, to check the impact of these different configurations. Each model was evaluated using the following prompt:

> *You are an AI medical assistant specializing in hypertension. Provide detailed and evidence-based answers, using clear and accessible language. Always respect patient privacy, and if you are unsure of the answer, state "I am not sure of the answer." Base your response on the provided context to answer accurately. Include current recommendations and explain medical concepts in an understandable way.*
> ***Context:*** { *context* }
> ***Question:*** { *question* }

Where { context } is the context provided to the model, and { question } is the question generated by the model. In case of RAG, the context is the retrieved information from the knowledge base, otherwise, it will be an empty string.

4.2. Metrics

To assess the performance of the different RAG systems, we employed a set of metrics specifically chosen to evaluate the retrieval and generation aspects of these systems. Following the RAGAS framework, we focused on three key metrics: *Answer relevancy* for assessing the relevance of the generated response to the original question; *Answer correctness* for assessing the factual correctness of the generated response; *Faithfulness* for evaluating the alignment of the generated response with the information present in the retrieved context. Each metric uses an LLM as a reference model to analyse the quality of the generated responses. For this study, we used GPT-4o as the reference model as it is one of the most advanced LLMs currently available.

Answer relevancy. This metric is computed by measuring the cosine similarity between the embedding of the original question, denoted as $\mathbf{E}_o$, and the embeddings of N generated questions, denoted as $\mathbf{E}_{g_i}$ where $i \in \{1, ..., N\}$. Formally:

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{E}_{g_i}, \mathbf{E}_o) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{E}_{g_i}\cdot\mathbf{E}_o}{||\mathbf{E}_{g_i}||||\mathbf{E}_o||}$$

where:

- $\cos(\mathbf{E}_{g_i}, \mathbf{E}_o)$ is the cosine similarity between the embedding of the i-th generated question and the embedding of the original question.
- $\mathbf{E}_{g_i}$ is the embedding of the i-th generated question.
- $\mathbf{E}_o$ is the embedding of the original question.
- N is the number of generated questions.

Note that while the cosine similarity ranges from -1 to 1, in practice, the answer relevance score typically falls between 0 and 1. This metric provides a measure of how well the generated responses align with the original question, with higher values indicating greater relevance.

Answer correctness. This metric, denoted as AC, measures the accuracy of a generated answer A with respect to a ground truth answer G. It combines two key aspects: factual correctness (FC) and semantic similarity (SS), both ranging from 0 to 1, with higher values indicating greater accuracy. In RAGAS, FC is computed using a language model to quantify the factual overlap between A and G: True Positive (TP): Statements present in both A and G; False Positive (FP): Statements present in A but not in G; False Negative (FN): Statements present in G but not in A. The F_1 score, a harmonic mean of precision and recall, is used to calculate FC:

$$FC = F_1 \text{ Score} = \frac{|TP|}{|TP| + 0.5 \times (|FP| + |FN|)}$$

For SS, RAGAS employs a language model to measure the semantic resemblance between A and G. The model generates embeddings for both answers, and their cosine similarity is calculated, resulting in the SS score. Finally, AC is computed as a weighted average of FC and SS:

$$AC = w_1 \times FC + w_2 \times SS$$

where w_1 and w_2 are user-defined weights that determine the relative importance of factual correctness and semantic similarity, respectively, with $w_1 + w_2 = 1$.

Faithfulness Let A be a generated answer and C be the given context. Let c_i represent a claim, where a claim is defined as a unit of information that can be independently verified. Define the set of claims in A as $C_A = \{c_1, c_2, ..., c_n\}$, where $|C_A| = n$. The faithfulness score, F, is calculated as follows:

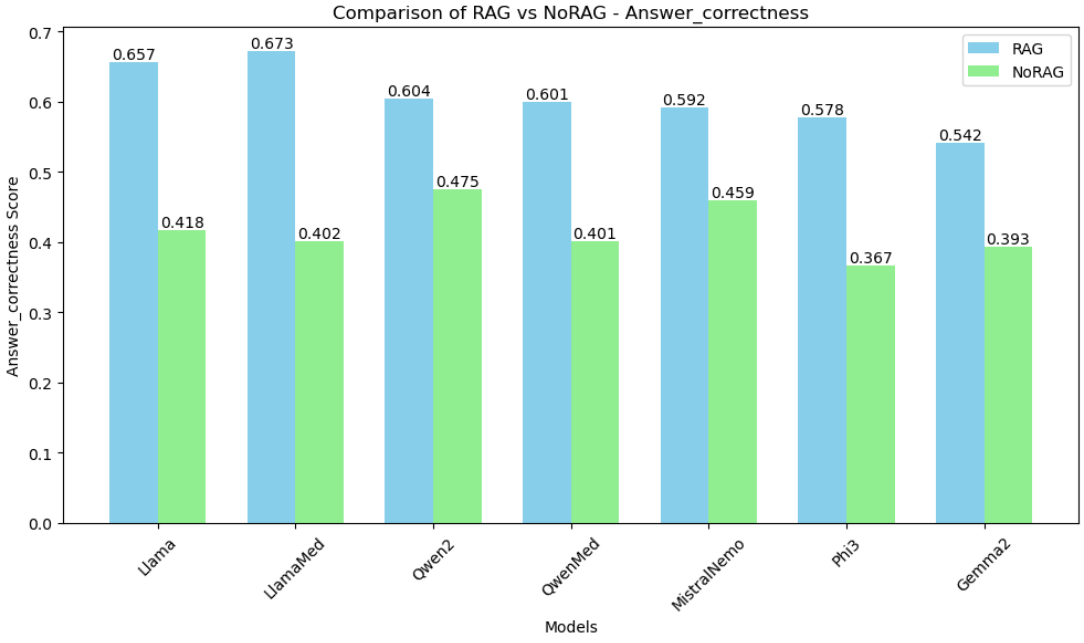
$$F(A, C) = \frac{|\{c_i \in C_A | c_i \text{ can be inferred from } C\}|}{n} \tag{1}$$

where $F(A, C) \in [0, 1]$. A higher score indicates greater factual consistency with the provided context.

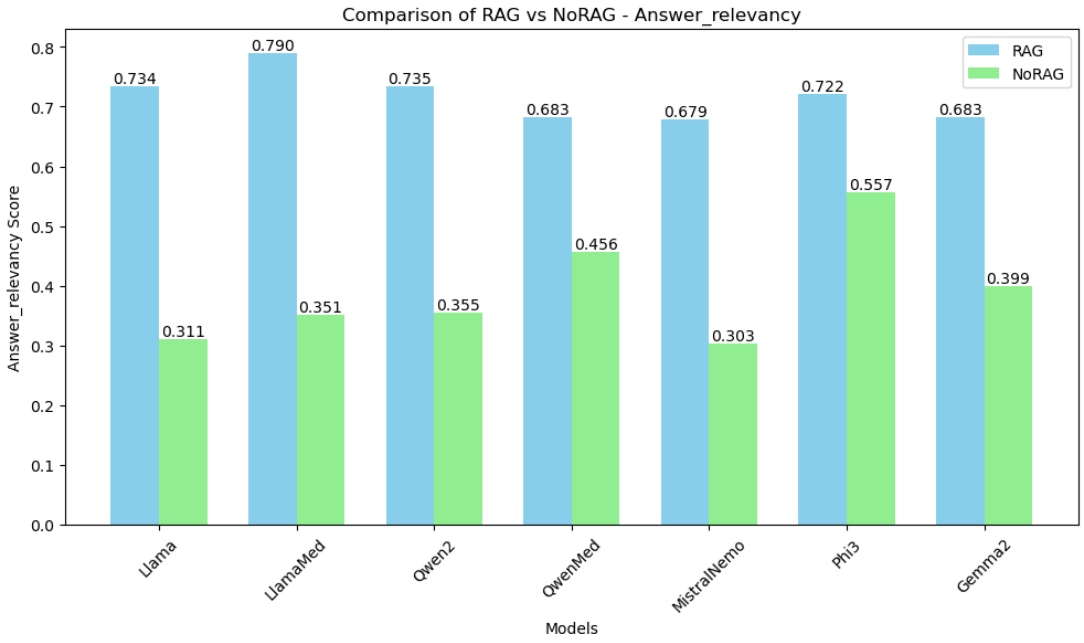
4.3. Results

This section presents the findings of our experiments, focusing on the performance comparison of RAG-based systems against plain LLM systems. We investigate the impact of different retrieval strategies and the effectiveness of domain-specific fine-tuning.

RAG vs. Base LLM Our experiments revealed that RAG-based systems consistently outperformed their plain LLM counterparts in terms of both answer relevancy and answer correctness (Figure 2). This observation strongly suggests that integrating external knowledge sources via retrieval augmentation significantly improves the quality of generated responses. The performance boost (up to 20% in answer correctness and 40% in answer relevancy in some cases) provided by RAG proved particularly relevant for models trained on medical data, showcasing the effectiveness of this approach in domain-specific tasks.



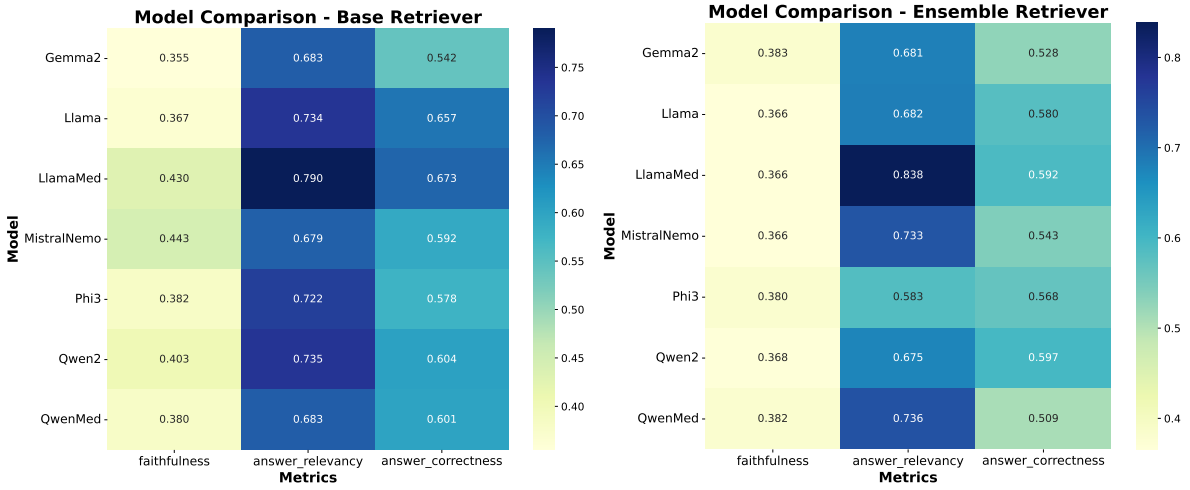
(a) Answer Correctness



(b) Answer Relevancy

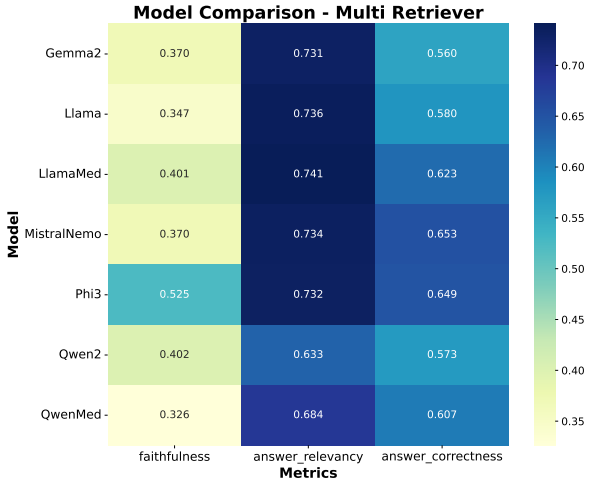
Figure 2: Performance comparison of RAG-based systems against plain LLM systems. (a) Answer Correctness, (b) Answer Relevancy. The results demonstrate that RAG significantly enhances both answer correctness and relevancy across all tested models.

Impact of Retrieval Strategies The choice of retrieval strategy significantly influenced the performance of the RAG systems. As depicted in Figure 3, each strategy exhibits strengths and weaknesses: Base Retriever excelled in answer correctness, relevance, and faithfulness, showcasing a good balance across all metrics; Ensemble Retriever demonstrated superior answer relevance but lagged in answer correctness and faithfulness; MultiQuery Retriever showed promise in answer relevance and faithfulness but suffered in answer correctness, potentially due to the increased likelihood of incorporating irrelevant information from the larger context. These results underscore the importance of aligning the retrieval strategy with the specific task and model. For instance, LlamaMed performed best with the



(a) Base Retriever

(b) Ensemble Retriever



(c) MultiQuery Retriever

Figure 3: Performance comparison of different retrieval strategies: (a) Base Retriever, (b) Ensemble Retriever, (c) MultiQuery Retriever. Each strategy demonstrates strengths and weaknesses across various evaluation metrics, highlighting the need for careful selection based on the specific task and model.

Base Retriever, while Mistral Nemo excelled with the MultiQuery Retriever. A detailed metric-based analysis is crucial for identifying the optimal strategy for a given scenario.

Domain-Specific Fine-tuning The comparison between base LLMs and specialised LLMs highlighted the benefits of domain-specific fine-tuning. Specialised models, such as LlamaMed and Qwen2 Med, consistently outperformed their base counterparts (Llama and Qwen2) in terms of answer relevance (up to 5%) and correctness (up to 3%). This finding underscores the value of tailoring LLMs to specific domains, particularly in specialised fields like healthcare, to maximise accuracy and relevance in generated responses.

RAG's Advantage Over specialised LLMs Our results showed that RAG-augmented base models consistently outperformed even the specialised LLMs without RAG across all evaluation metrics (see Figure 2). This observation suggests that when sufficient data is available to construct a comprehensive knowledge base, employing RAG techniques can be more advantageous than solely relying on domain-specific fine-tuning. This finding holds significant practical implications, particularly in scenarios where data scarcity might make fine-tuning challenging or impractical.

4.4. Discussion

The findings from our study underscore the effectiveness of RAG as a robust tool for enhancing the capabilities of open-source LLMs in the medical domain. By incorporating a knowledge base of verified medical data, RAG significantly improves both the accuracy and relevance of the chatbot's responses, making it an optimal solution for supporting hypertensive patients.

One key insight from the study is that RAG-augmented models consistently outperform specialised LLMs. This suggests that even without large datasets for fine-tuning, RAG can serve as an efficient, out-of-the-box solution for creating high-performance chatbots in specialised domains. Furthermore, our experiments emphasize the critical importance of choosing an appropriate retrieval strategy tailored to the specific task and model. A thorough evaluation of different retrieval methods is essential to ensure optimal performance before deployment.

A limitation of this study lies in the evaluation methodology. We utilised the RAGAS framework, which assesses chatbot responses based on a set of metrics using a reference model. However, given that the reference model is a LLM itself, it may not capture all aspects of medical knowledge and context. To ensure the quality of the responses, therefore, for future work, we suggest incorporating human evaluation by medical professionals to assess the clinical accuracy and appropriateness of the chatbot's responses.

Another aspect not addressed in our study is the measurement of empathetic responses. While the generated replies were accurate and relevant with respect to the provided dataset (which was reviewed by medical professionals), we did not directly evaluate the empathy of the responses, which is crucial in patient interactions. Future research should focus on establishing metrics for assessing empathy, potentially involving a group of experts to evaluate responses from this perspective.

5. Conclusion

This study investigated the application of RAG on open-source LLMs for developing a medical empathetic chatbot designed to support hypertensive patients. Driven by the need for data privacy, accuracy, and empathetic communication, we explored RAG as a viable alternative to proprietary LLMs and fine-tuning approaches. Our experiments evaluated several open-source LLMs, including both general-purpose and medical domain-specific models, across three different retrieval strategies: Base Retriever, Ensemble Retriever, and MultiQuery Retriever.

Our findings demonstrate that incorporating RAG significantly improves the performance of open-source LLMs in this specific medical context. Across all tested models, RAG consistently enhanced both answer relevance and correctness compared to their base LLM counterparts. Notably, RAG-augmented models outperformed even the specialised, medically fine-tuned LLMs, indicating the potential of RAG as a powerful and efficient alternative, particularly in scenarios where fine-tuning data is limited.

However, this study has limitations. The evaluation relied on the RAGAS framework, which utilizes another LLM (GPT-4) as a reference model. While this approach offers a standardised and comprehensive evaluation, it might not fully capture the complexities of medical expertise and context. Furthermore, our evaluation did not explicitly address the empathy of generated responses, a crucial aspect of patient interaction in healthcare applications.

Future research should focus on two key areas: First, incorporating human evaluation by medical professionals to assess the clinical accuracy and appropriateness of the chatbot's responses. Second, developing and integrating metrics specifically designed to evaluate the empathy and emotional intelligence of the chatbot's communication. Addressing these limitations will pave the way for developing more robust and clinically valuable empathetic chatbots for supporting patients in managing chronic conditions like hypertension.

Acknowledgments

This work has been partially funded by the European Union - NextGenerationEU within the framework of PNRR Mission 4 - Component 2 - Investment 1.1 under the Italian Ministry of University and Research (MUR) programme "PRIN 2022" - grant number 2022N2NH42 - SmartShires - CUP: H53D23003570006

References

[1] S. Montagna, G. Aguzzi, S. Ferretti, M. F. Pengo, L. C. Klopfenstein, M. Ungolo, M. Magnini, LLM-based Solutions for Healthcare Chatbots: a Comparative Analysis, in: 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 2024, pp. 346–351. doi:10.1109/PerComWorkshops59983.2024.10503257.

[2] S. Montagna, S. Ferretti, L. C. Klopfenstein, A. Florio, M. F. Pengo, Data decentralisation of llm-based chatbot systems in chronic disease self-management, in: Proceedings of the 2023 ACM Conference on Information Technology for Social Good, GoodIT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 205–212. doi:10.1145/3582515.3609536.

[3] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, D. M. Smith, Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum., JAMA Internal Medicine (2023). doi:10.1001/jamainternmed.2023.1838.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[5] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, S. J. Wagner, J. N. Kather, The future landscape of large language models in medicine, Communications Medicine 3 (2023) 141. doi:10.1038/s43856-023-00370-1.

[6] H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, J. W. Gichoya, Ethics of large language models in medicine and medical research, The Lancet Digital Health 5 (2023) e333–e335. doi:10.1016/S2589-7500(23)00083-3.

[7] J. Haltaufderheide, R. Ranisch, The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms), npj Digital Medicine 7 (2024) 183. doi:10.1038/s41746-024-01157-x.

[8] M. Giuffrè, S. Kresevic, N. Pugliese, K. You, D. L. Shung, Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes, Liver International 44 (2024) 2114–2124. doi:https://doi.org/10.1111/liv.15974.

[9] S. Kresevic, M. Giuffrè, M. Ajcevic, A. Accardo, L. S. Crocè, D. L. Shung, Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework, npj Digital Medicine 7 (2024) 102. URL: https://doi.org/10.1038/s41746-024-01091-y.

[10] J. Maharjan, A. Garikipati, N. P. Singh, L. Cyrus, M. Sharma, M. Ciobanu, G. Barnes, R. Thapa, Q. Mao, R. Das, Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models, Scientific Reports 14 (2024) 14156. doi:10.1038/s41598-024-64827-6.

[11] A. Wang, C. Liu, J. Yang, C. Weng, Fine-tuning large language models for rare disease concept normalization, Journal of the American Medical Informatics Association 31 (2024) 2076–2083. doi:10.1093/jamia/ocae133.

[12] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models:

A comprehensive survey, CoRR abs/2403.14608 (2024). doi:10.48550/ARXIV.2403.14608. arXiv:2403.14608.

[13] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

[16] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, 2nd Ed, Cambridge University Press, 2014. URL: http://www.mmds.org/.

[17] S. E. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. doi:10.1561/1500000019.

[18] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.

[19] J. G. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, 1998, pp. 335–336. doi:10.1145/290941.291025.

[20] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, CoRR abs/2402.01613 (2024). doi:10.48550/ARXIV.2402.01613. arXiv:2402.01613.

[21] H. Chase, LangChain, 2022. URL: https://github.com/langchain-ai/langchain.

[22] A. D. et al., The Llama 3 Herd of Models, CoRR abs/2407.21783 (2024). doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[23] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[24] A. Dubey, et. al., The Llama 3 Herd of Models, 2024. doi:10.48550/ARXIV.2407.21783.

[25] A. Yang, et. al., Qwen2 technical report, 2024. doi:10.48550/ARXIV.2407.10671.

[26] M. Abdin, et. al., Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL: https://arxiv.org/abs/2404.14219. arXiv:2404.14219.

[27] G. Team, Gemma 2: Improving open language models at a practical size, 2024. URL: https://arxiv.org/abs/2408.00118. arXiv:2408.00118.