UNIVERSIDADE FEDERAL DE PERNAMBUCO

CHARLES GABRIEL CARVALHO CRISTOVÃO

Comparative Benchmarking of Retrieval-Augmented Generation Reranker for Medical Domain

Recife
2025

CHARLES GABRIEL CARVALHO CRISTOVÃO

Comparative Benchmarking of Retrieval-Augmented Generation Reranker for Medical Domain

Relatório final, apresentado a Universidade Federal de Pernambuco, como parte das exigências para a obtenção do título de bacharel em ciência da computação.

Orientador: Prof. Tsang Ing Ren

Recife
2025

CHARLES GABRIEL CARVALHO CRISTOVÃO

Comparative Benchmarking of Retrieval-Augmented Generation Reranker for Medical Domain

Relatório final, apresentado a Universidade Federal de Pernambuco, como parte das exigências para a obtenção do título de bacharel em ciência da computação.

Recife, 03 de abril de 2025.

BANCA EXAMINADORA

_____

Prof. (Tsang Ing Ren)
Universidade Federal de Pernambuco (UFPE), Centro de Informática (CIn)

_____

Prof. (Ricardo Bastos Cavalcante Prudencio)
Universidade Federal de Pernambuco (UFPE), Centro de Informática (CIn)

Comparative Benchmarking of Retrieval-Augmented Generation Reranker for Medical Domain

Charles G. C. Cristovão Tsang Ing Ren

[1]Centro de Informática
Universidade Federal de Pernambuco (UFPE) – Recife, PE – Brazil

{cgcc, tir}@cin.ufpe.br

Abstract. *The exponential growth of digital medical information poses significant challenges in delivering reliable, evidence-based responses to clinical inquiries. Traditional systems often fall short in bridging the gap between vast data repositories and the need for authoritative, contextually relevant insights. In this study, we introduce a pipeline that leverages a Retrieval-Augmented Generation reranker architecture, combined with a Chain-of-Thought (CoT) prompting strategy, to enhance the performance of Large Language Models in addressing complex medical questions. By integrating a robust retrieval mechanism that sources trustworthy evidence from established medical literature and by refining the information with reranking, our approach not only improves answer accuracy but also demonstrates that larger models can be effectively distilled into smaller, more resource-efficient variants while maintaining comparable performance. The pipeline is evaluated in zero-shot question-answering scenarios, employing a question-only retrieval strategy to simulate realistic clinical contexts where prior domain-specific fine-tuning is absent. This work underscores the potential of combining retrieval techniques with sequential reasoning to overcome the inherent challenges in medical AI, paving the way for more accurate, transparent, and accessible systems in healthcare applications.*

1. Introduction

With the advancement of Large Language Models (LLM's), significant improvements have been made across various domains, including the medical field. These models possess broad general knowledge, enabling them to answer questions without prior context. However, they still exhibit limitations in generating well-founded responses and interpreting complex medical terminology within clinical texts. These shortcomings underscore the need to explore approaches that enhance the reliability and precision of LLM-generated responses in this context.

Recent studies have shown that integrating information retrieval mechanisms can mitigate some of these limitations. Models that combine Retrieval-Augmented Generation (RAG) techniques have proven effective in extracting relevant evidence from medical literature, thereby providing more reliable and well-supported responses [9]. This approach not only enhances the accuracy of the answers but also enables smaller, less computationally demanding models to achieve performance comparable to larger models, making them more accessible and efficient.

In addition to retrieval-based approaches, advances in prompting techniques have shown significant potential to improve LLM reasoning for medical applications. Chain-of-Thought reasoning [6], for example, allows models to follow an incremental reasoning

process, simulating clinical decision-making more effectively. This technique enhances response interpretability and enables more accurate differential diagnoses by iteratively refining medical reasoning. Moreover, rerankers mechanisms have been introduced to further enhance reliability. By leveraging reward models trained on expert-validated medical question-answer pairs, these approaches can filter out incorrect or misleading responses, ensuring higher trustworthiness in clinical applications. This verification step is crucial in medical AI, where incorrect information can lead to severe consequences in decision-making.

This study examines the impact of incorporating RAG techniques, rerankers, and advanced prompting methods, such as Chain-of-Thought, on improving the accuracy and reliability of LLMs in addressing complex medical questions. We assess how these approaches enhance model performance and contribute to generating trustworthy responses.

Our goal is to contribute to the development of scalable and efficient solutions that can be applied in real-world healthcare scenarios, thereby expanding the practical utility of LLMs in clinical practice. By discussing the implementation and evaluation of these techniques, we explore their impact on response quality and computational efficiency. We believe that this research will contribute to the advancement of LLMs in medicine, enabling the creation of more accurate, transparent, and accessible systems for healthcare professionals and researchers.

2. Related Work

One of the pioneering works that formalized the concept of RAG is the seminal paper by Patrick Lewis [3], titled "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In this groundbreaking study, the authors proposed integrating a document retrieval system with a generative language model. This innovative framework enables the model to incorporate relevant, up-to-date external information into its responses, thus mitigating issues such as hallucinations and outdated knowledge that are inherent in models trained on static datasets. The work laid a strong theoretical and practical foundation, influencing many subsequent advances in RAG techniques.

Beyond RAG, the broader field of open-domain question answering (OpenQA) has played a crucial role in shaping retrieval-based methodologies. Traditional QA tasks often relied on predefined, human-curated texts as context for answering questions, a practice that is impractical for real-world applications due to the high cost of manual annotation. This challenge led to the emergence of OpenQA, where models must autonomously retrieve and comprehend relevant information to generate accurate responses. Early efforts in this direction, such as the DrQA model [2] employed a two-step approach: first, retrieving relevant documents from a large knowledge base like Wikipedia, and then applying a reading comprehension model to extract the most relevant answer. Subsequent research refined these methods by introducing more sophisticated retrieval mechanisms, either aggregating multiple sources of evidence or filtering out less relevant content to improve answer accuracy. These advancements laid the groundwork for integrating retrieval-based techniques with modern LLMs.

Recent advancements have introduced CoT prompting to further enhance the reasoning capabilities of LLM's in medical contexts. CoT prompting encourages models to generate intermediate reasoning steps that mirror human thought processes, resulting

in more accurate and interpretable responses. For example, one study showed that providing models with just two examples of diagnostic reasoning using CoT led to a 15% increase in diagnostic accuracy compared to traditional methods [5]. In another study, CoT strategies were applied to detect and correct medical errors, demonstrating significant improvements in the precision and reliability of language models in the healthcare domain.[8].

Chen et al. (2025) [1] emphasize the feasibility of utilizing smaller models to tackle complex problems, achieving performance levels comparable to larger models. By introducing MedOmniKB—a repository encompassing diverse medical knowledge sources—and the Source Planning Optimization (SPO) method, the study demonstrates that optimized smaller models can effectively plan and align with various knowledge sources. Experimental results confirm that these optimized smaller models outperform existing methods in multi-source planning, efficiently leveraging diverse medical knowledge sources.

Building on these foundations, new frameworks have emerged specifically tailored for the medical domain. Notably, the MedRAG framework [9] integrates domain-specific retrieval with generative processes, achieving significant improvements in handling open-ended medical questions. Its iterative variant, iMedRAG, further refines responses by incorporating multi-step reasoning that mirrors clinical decision-making. Additionally, PMC LLaMA—a variant of the LLaMA model fine-tuned on extensive biomedical corpora from PubMed Central—has demonstrated promising performance in extracting and generating clinically relevant content [7]. Collectively, these studies underscore the potential of combining advanced prompting strategies with retrieval and domain adaptation to produce more robust and interpretable outputs.

3. Proposed Pipeline and Methodology

The pipeline proposed in this study was developed by integrating a RAG reranker with a Chain of Thought to tackle multiple-choice questions in the medical domain. It leverages segmented textbook content and curated snippets to assemble a diverse evidence base. First, semantic embedding techniques are employed to retrieve relevant text segments from a comprehensive repository of medical resources. Then, reranker—guided by CoT prompting—prioritizes and refines these segments based on their relevance and coherence. The final stage synthesizes the curated information into a structured answer that addresses complex medical queries efficiently.

3.1. Corpora

The materials used in the RAG pipeline consist of a combination of textbooks provided by the MedQA dataset [2] and snippets employed in MedRAG [9]. In total, 18 textbooks were incorporated, which are widely used in preparing students for the USMLE, the mandatory examination for obtaining a medical license in the United States. The textbooks are divided into smaller segments using a semantic splitter that preserves semantically related sentences. The snippets utilized in medRag were generated from articles available on PMC; however, only a small fraction of these snippets—totaling 100,000—were employed in the pipeline.

The snippets are structured as follows:

- **Id:** Unique identifier.
- **Title:** The title of the PubMed article from which the snippet was extracted.
- **Content:** The abstract of the article.
- **Contents:** A concatenation of the title and the abstract.
- **PMID:** The identifier of the article in PMC.

3.2. Models

A total of four state-of-the-art models with varying parameter sizes were employed to assess the impact of model scale on the results. Below, we describe each model in detail, highlighting their key characteristics and how they contribute to either text processing or reasoning capabilities:

- **LLaMA 3.2:** This model, featuring 3 billion parameters, is designed for efficient text processing. Its compact size makes it ideal for tasks that require quick inference and lower computational overhead while still delivering robust performance in natural language understanding.

- **Mistral:** With 7 billion parameters, Mistral strikes a balance between efficiency and capability. It offers enhanced performance over smaller models, particularly in more complex text processing tasks, while maintaining a relatively low computational footprint.

- **DeepSeek R1:** This 8-billion-parameter model is tailored for reasoning tasks, providing improved logical inference and analytical abilities. DeepSeek R1 demonstrates a strong capacity for complex problem-solving and deeper contextual understanding, making it particularly effective for tasks requiring reasoning over text.

- **Phi4:** The largest of the group, with 14 billion parameters, delivers state-of-the-art performance in both text processing and reasoning. Its extensive parameter count allows for nuanced understanding and sophisticated reasoning capabilities, which significantly enhance its performance in tasks that demand a high level of inference and context comprehension.

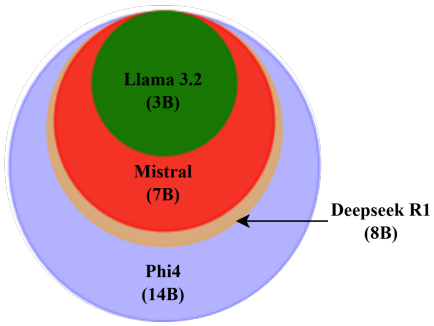

Figure 1. Model sizes comparison

Each model was chosen to explore the trade-offs between model size, efficiency, and performance, providing a comprehensive evaluation of how parameter scale influences the effectiveness of the proposed pipeline.

3.3. Pipeline

The proposed pipeline combines a Retrieval-Augmented Generation model with a reranker and a Chain of Thought method to enhance the model's ability to accurately answer multiple-choice medical questions. The embedding model used is all-MiniLM-L6-v2, selected to maintain a balance between the time required to generate vector representations of the reference material and efficiency in semantic context tasks. This section details the pipeline structure, processing flow, and the integration of CoT in response generation.

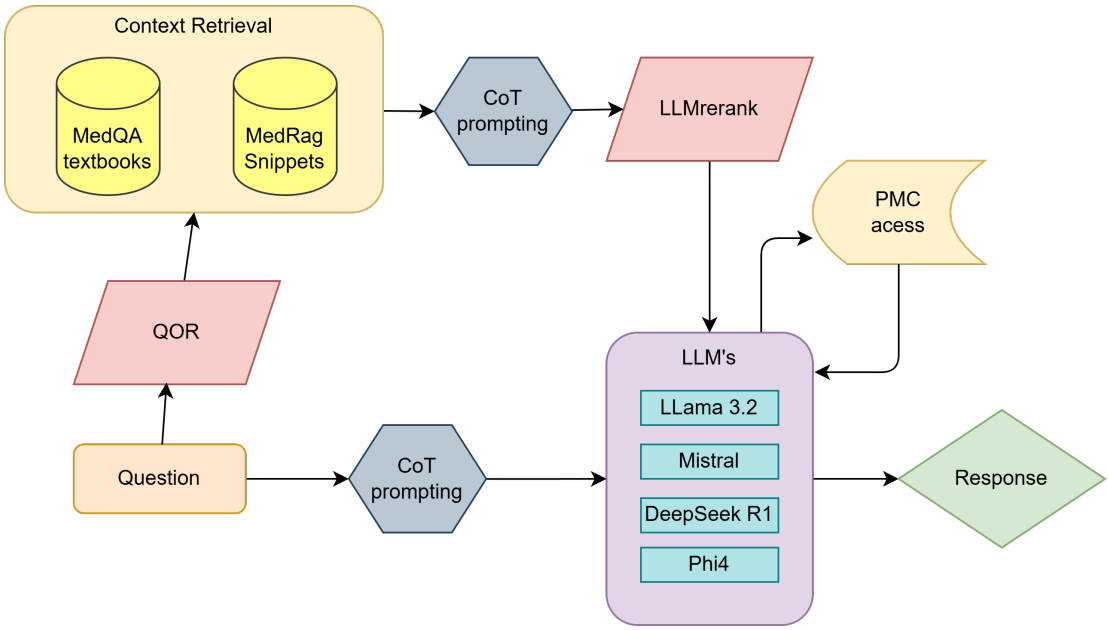

Figure 2. This diagram illustrates a pipeline where a question undergoes retrieval, CoT prompting, and LLM reranking before being processed by a chosen LLM (LlaMA 3.2, Mistral, DeepSeek R1, or Phi4). The system may also access PMC for additional medical information before generating the final response.

The pipeline follows a three-step process:

1. Context Retrieval:
   The system initiates its response process by conducting an in-depth search within reference materials, including segmented textbooks and medRag snippets, to locate evidence and information that will underpin the provided answer.
   To achieve this goal, the information retrieval component employs semantic embedding. These techniques transform the content of the reference materials into representations that allow for more efficient and relevant searches. This approach enables the system to accurately identify the most pertinent excerpts within a knowledge base composed of segmented textbooks and medRag snippets.
   Furthermore, to simulate real-world search scenarios, the system adopts the "question-only retrieval" strategy proposed by Xiong et al [9]. In this approach, the search is conducted using only the question as the criterion, without considering any answer options during the retrieval phase. This simulates practical sit-

uations where a user has only the question at hand and needs to find relevant information without the aid of predefined answer choices.
This method allows the system to be both flexible and effective in retrieving data, ensuring that the extracted information directly relates to the content of the question, ultimately enhancing the quality and accuracy of the final response.

2. Evidence Reranking:
Once the initial set of excerpts is retrieved, the system employs a reranker to further refine and prioritize this evidence. Implemented using the Llama Index, the reranker is guided by a CoT prompt that instructs it to evaluate the relevance and coherence of each fragment. This prompt-based approach enables the reranker to identify which content is most critical for answering the question.
To ensure robust evaluation, the reranker processes the retrieved documents in batches. For instance, it compares and ranks groups of documents—such as determining the relative importance among five documents at a time—thereby systematically prioritizing the most useful information. This method ensures that the final response is underpinned by the most pertinent evidence, enhancing both the quality and accuracy of the answer provided.

3. Response Generation:
The response generation stage integrates the question with the reranked text segments to produce an answer grounded in the most relevant evidence. During this phase, Chain of Thought prompting is employed to enhance reasoning and provide a detailed explanation. This approach is particularly beneficial given that the dataset comprises two distinct types of questions:
Type 1: These questions require a single reasoning step and focus on one specific piece of knowledge. For example, a question like "Which of the following symptoms is associated with schizophrenia?" [2] demands a straightforward answer based on a single fact.
Type 2: These questions simulate realistic clinical scenarios by describing a patient's condition and then asking for the most likely diagnosis, appropriate treatment, necessary examination, underlying mechanisms of certain conditions, or potential outcomes of a given treatment [2]. Such questions are more complex and necessitate multi-step reasoning. For instance, a question detailing a 27-year-old man with painful urination requires the model to extract and interpret symptoms to infer the underlying cause.
By utilizing CoT prompting, the generation model is guided to break down these complex questions into intermediate reasoning steps. This structured thought process enables the model to accurately integrate evidence, handle both simple and multifaceted queries, and ultimately produce a coherent and well-supported answer.
Additionally, during the answer generation phase, if the retrieved excerpts provide insufficient information, the LLM is further guided to seek additional data from the full articles. This is often possible when the retrieved text segments include metadata, such as a PMID, which can be used to locate and reference the complete article on the PMC website. This strategy ensures that the response is comprehensive and backed by complete, authoritative sources when necessary.

3.4. Tools and implementation

The pipeline was implemented using robust frameworks to ensure efficient processing and integration of large language models. The key tools utilized include:

- **LlamaIndex:** LlamaIndex was employed to structure and manage data within the pipeline, facilitating efficient integration with LLMs and enabling effective retrieval and manipulation of relevant information.
- **Ollama:** Ollama was utilized to process embeddings and manage the local execution of models.
- **Instructor:** It was employed to ensure that the model outputs were structured according to the specific requirements of the tests, facilitating the analysis and interpretation of results [4].

4. Experiments and results

The experiments were conducted using the MedQA dataset, which consists of 1273 multiple-choice questions from the medical domain, each offering five answer options. The primary objective was to evaluate the efficacy of the proposed pipeline and to determine whether incorporating the retrieval process during inference leads to a measurable improvement in model performance.

In these experiments, the four state-of-the-art models were tested under two distinct configurations: one utilizing only the Retrieval-Augmented Generation approach and the other combining RAG with the Chain of Thought technique. The RAG configuration leverages a question-only retrieval mechanism to simulate realistic scenarios where only the query is available, while the RAG + CoT configuration integrates a sequential reasoning process designed to enhance the model's ability to justify its conclusions.

Performance was assessed in a zero-shot setting, focusing on each model's ability to correctly determine the answer without prior fine-tuning on the task-specific data. In addition to evaluating the absolute performance of these configurations, the experimental results were compared with benchmark data obtained from previously tested models on the MedQA dataset.

Model	Baseline	RAG	RAG+CoT
Phi4	73.61%	74.31%	78.08%
Mistral	41.95%	39.75%	52.63%
LLaMA 3.2	51.14%	52.32%	54.60%
DeepSeek-R1	39.91%	45.48%	39.98%

Table 1. Baseline model and variants accuracy

For instance, the considerable performance jump observed in the Mistral model implies that its primary limitation lies in the reasoning process rather than in the retrieval mechanism or underlying language modeling capabilities. Conversely, the moderate gains seen in models like LLaMA 3.2 indicate that while these models are already robust, the incorporation of structured reasoning still confers a measurable advantage. Notably, the

contrasting behavior of DeepSeek-R1—which did not exhibit further improvement with chain-of-thought despite benefiting from retrieval alone—underscores the variability in how different architectures leverage supplementary reasoning techniques.

Overall, these nuanced results emphasize the importance of aligning model design with the specific challenges posed by the task and advocate for further exploration of customized chain-of-thought approaches tailored to individual model architectures.

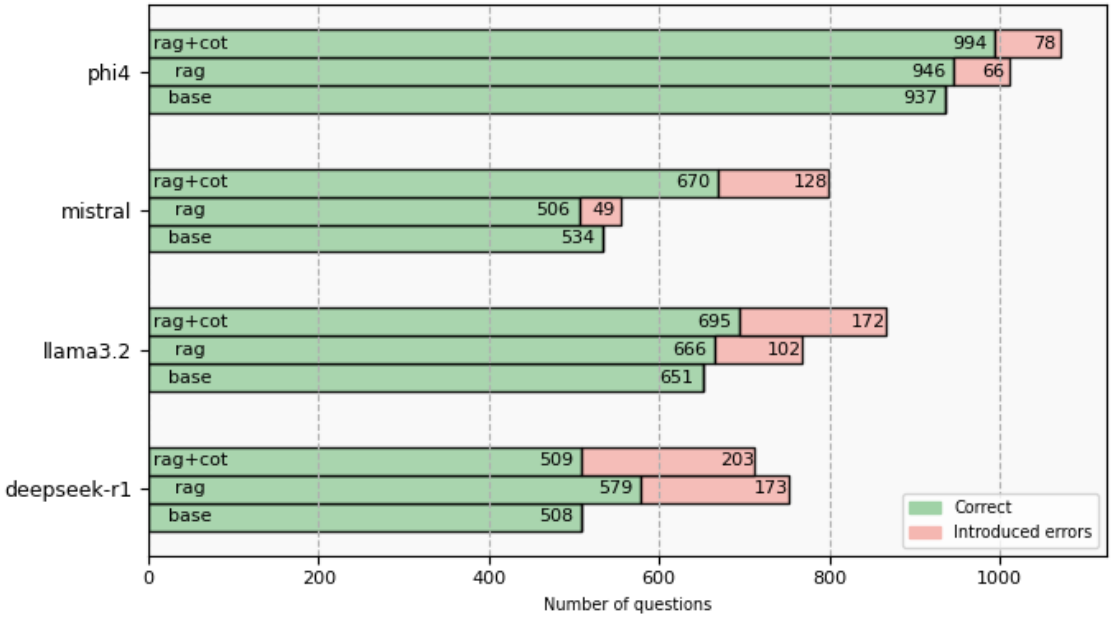

Figure 3. Variant introduced errors

The analysis of errors introduced by the variants reveals a nuanced trade-off between increasing correct responses and the inadvertent generation of new errors. For example, the DeepSeek-R1 model saw an increase from 508 to 579 correct answers with the RAG variant, yet this improvement came at the cost of 173 new errors, and the RAG+CoT variant further exacerbated this issue, yielding 203 new errors with only a marginal return to baseline correct answers. In contrast, LLaMA 3.2 and Phi4 experienced moderate increases in correct answers—with LLaMA 3.2 rising from 651 to 695 and Phi4 from 937 to 994—while the number of new errors remained relatively contained. The Mistral model displayed an interesting pattern: although the RAG variant reduced correct answers from 534 to 506 and introduced 49 new errors, the subsequent integration of chain-of-thought reasoning recovered performance dramatically to 670 correct answers, albeit with an increased error count of 128.

These results suggest that while retrieval-augmented techniques and structured reasoning can enhance model performance, they also introduce additional errors, likely due to noise in the retrieval process or errors in the reasoning chain. This trade-off highlights the need for further optimization to balance the benefits of increased accuracy with the minimization of error propagation.

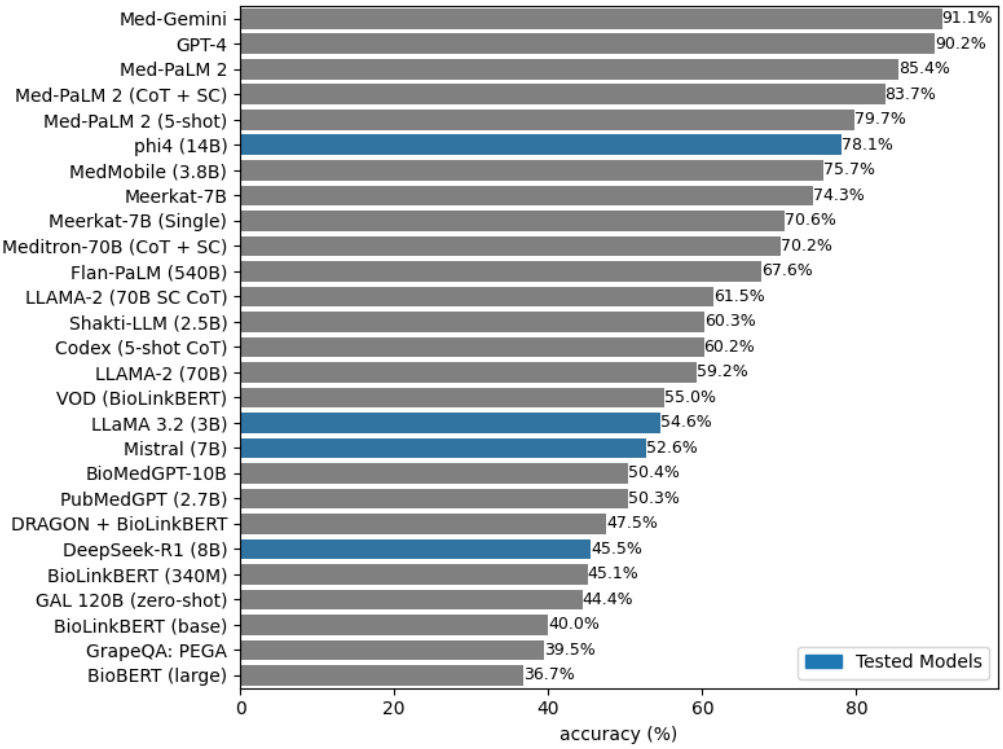

Figure 4. Comparison benchmark

Despite having a relatively limited number of parameters, the models performed well on the MedQA benchmark, particularly phi4. This outcome demonstrates that recent, more optimized architectures can achieve performance levels close to state-of-the-art models, even when scaled down, highlighting the effectiveness of modern design strategies in bridging the gap between efficiency and high performance.

5. Conclusion

This study set out to investigate the efficacy of a Retrieval-Augmented Generation pipeline enhanced by the Chain of Thought technique for answering complex medical multiple-choice questions. The objective was to address the inherent challenges of zero-shot reasoning in the medical domain, a task complicated by the need for precise and well-justified responses.

In summary, the proposed pipeline has demonstrated an enhancement in model performance when tackling complex medical multiple-choice questions in a zero-shot setting. By combining state-of-the-art retrieval mechanisms with sequential reasoning, the approach not only improves the models' ability to correctly identify answers but also bridges the gap between raw information retrieval and informed decision-making. This integrated method provides a clear advancement over traditional approaches, underscoring the potential of structured reasoning in complex domains.

However, these performance gains come with notable trade-offs. While the pipeline generally leads to an increase in correct responses, it also introduces additional errors that raise concerns about the noise generated during the retrieval and reasoning stages. This drawback suggests that the benefits of the combined RAG+CoT approach

are sometimes offset by the propagation of errors, especially in models with limitations in their architectural capacity for sequential reasoning. Consequently, future research should focus on mitigating this noise—potential strategies include incorporating advanced computational techniques such as self-consistency checks, employing more sophisticated embedding models for improved semantic search, and refining reranking methods. Moreover, evaluating fine-tuned models could offer insights into whether a more optimized model configuration might achieve significant gains with reduced error rates.

An additional challenge encountered during this study was the scarcity of suitable datasets in the medical domain. Many similar datasets lack publicly available ground-truth labels, which not only complicates the testing process but also necessitates empirical adjustments during experimentation. Addressing this limitation by developing or accessing more comprehensive datasets would be instrumental in facilitating more rigorous evaluations and potentially broadening the applicability of the approach.

In conclusion, this work demonstrates the promising potential of a RAG pipeline augmented with CoT in advancing the capabilities of medical question-answering systems. The approach not only contributes to current academic discourse but also lays the groundwork for future research aimed at overcoming the identified challenges and broadening the applicability of such systems in diverse domains.

References

[1] Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications, 2025.

[2] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[4] Jason Liu and Contributors. Instructor: A library for structured outputs from large language models, 3 2024.

[5] Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1), 2024.

[6] Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering. *arXiv e-prints*, page arXiv:2403.04890, March 2024.

[7] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 04 2024.

[8] Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason Cheung, Teng Zhang, and Honghan Wu. KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-though (CoT) prompting strategies for medical error detection and correction. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman, editors, *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 353–359, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.