Medical Question Answering System Based on Retrieval-augmented Generation

Chengwei Cao[1*]

[1]School of Computer Sciences, University Sains Malaysia, Penang 11800, Malaysia

Abstract. This study investigates the feasibility and reliability of Retrieval-Augmented Generation (RAG) in medical question-answering tasks. To address issues such as hallucination and lack of traceability in large language models (LLMs) in medical contexts, a medical knowledge base was constructed using the cancer-related subset of a Kaggle healthcare dataset. Experiments were conducted with the Qwen-Plus and Qwen-Flash models. Evaluation was carried out across three dimensions: answer accuracy, source traceability, and refusal capability, with additional analysis on the impact of different retrieval quantities (top-k) on performance. The results show that RAG significantly improves the semantic consistency of model responses, outperforming the baseline model on the BERTScore F1 metric. It also demonstrates strong performance in terms of refusal rate and attribution accuracy, highlighting its advantages in mitigating hallucinations and enhancing interpretability. Furthermore, the findings indicate that a retrieval quantity of k=5 yields the best overall performance. This study validates the potential of RAG in medical question answering and provides empirical support for building trustworthy medical AI systems.

1 Introduction

Some emerging large language models (LLMs) demonstrate excellent capabilities in understanding and generating natural language, answering questions, producing text, and providing knowledge assistance. Models such as GPT and Gemini, trained on massive datasets with vast numbers of parameters, achieve impressive performance in these general tasks. However, their application in high-risk or highly specialized fields like medicine remains significantly limited. Only a portion of their outputs can be trusted for accuracy and reliability; they lack domain-specific expertise, their knowledge may become outdated, and it is often difficult to trace their answers back to authoritative sources [1]. These limitations pose substantial challenges and risks for using LLMs in the medical field. If deployed in clinical practice, LLMs could potentially generate incorrect and hazardous information, directly impacting clinical judgments and patient safety [2]. Therefore, investigating methods to leverage their generative capabilities while ensuring the provision of accurate, professional knowledge with traceable sources represents a critical research direction for both academia and industry.

* Corresponding author: cao041125@student.usm.my

To tackle these issues, researchers such as Patrick Lewis introduced the Retrieval-Augmented Generation (RAG) framework in a 2020 paper [3]. The core idea behind RAG is straightforward: before generating a response, the system retrieves relevant information from external authoritative knowledge sources. It then feeds both the retrieved documents and the user's query into the language model. By doing so, RAG combines the language generation ability of LLMs with up-to-date external knowledge, improving not only the accuracy and expertise of answers but also their explainability [4]. This approach has proven effective in reducing model hallucinations. As a result, RAG is increasingly being applied in highly sensitive and specialized fields like healthcare, supporting the development of more trustworthy medical AI systems [5].

Currently, research and application of RAG in the healthcare domain remain in the exploratory stage, and its potential is far from being fully realized. This paper will focus on the following three aspects: First, it examines the effectiveness of RAG in mitigating hallucination issues, particularly in medical contexts where the model should refrain from responding when uncertain rather than generating potentially misleading information. Second, it evaluates the traceability of knowledge sources in responses, verifying whether answers are grounded in retrieved medical literature or knowledge bases to enhance the credibility and interpretability of the system. Third, it systematically analyzes how the configuration of retrieval parameters—such as the top-k setting—affects answer quality, with special attention to potential information loss or redundancy caused by different retrieval scales. Through investigation along these three dimensions, this study aims to provide a comprehensive assessment of the applicability and limitations of RAG in medical question answering.

2 Related Work

In recent years, LLMs have shown growing potential in medical question answering and clinical decision support. Models like those in the GPT series, for example, have achieved promising results on tasks such as medical exam questions and case analyses [6]. Yet, multiple studies point out that LLMs continue to struggle with a critical problem in medical applications: the tendency to generate hallucinations. They often produce answers that lack factual support or even contradict basic medical knowledge [7]. These models also face limitations in specialization and traceability, largely due to constraints in the timeliness and depth of the training data they rely on. Such issues currently prevent their direct use in real clinical and healthcare environments.

To tackle the problem of hallucinations in LLMs, the RAG framework was introduced [3]. By incorporating a retrieval step that gathers information from external knowledge sources before generating a response, RAG allows models to ground their outputs in up-to-date and authoritative documents. Studies show that this method significantly improves not only the accuracy and expertise of model responses, but also their interpretability. RAG has proven effective in knowledge-intensive applications including open-domain QA and scientific summarization [8, 9]. According to relevant reviews, it is also considered a key strategy for mitigating hallucinations and enabling knowledge updates in LLMs [10].

In the medical field, research on RAG is attracting increasing attention. Previous studies have shown that RAG can effectively enhance both the quality and verifiability of responses in medical question-answering tasks. For example, the MedRAG system proposed by Xiong et al. systematically evaluated the performance of RAG in medical Q&A, demonstrating its ability to significantly improve model accuracy and interpretability [11]. Experiments conducted by Shi et al. with MKRAG also indicated that incorporating external knowledge can enhance RAG's performance in medical question answering, even without retraining the model [12]. These findings lay the groundwork for the practical application of RAG in

medical intelligent systems. This study examines three key dimensions: hallucination mitigation (non-response capability), knowledge attribution (source verification rate), and retrieval parameter configuration (top-k settings). First, it evaluates the model's ability to refuse to answer under uncertain conditions to avoid generating misleading information in medical contexts. Second, it assesses whether answers are grounded in retrieved medical literature and whether they are traceable and interpretable. Third, it investigates how different retrieval scales affect the quality and reliability of responses. Through systematic experimentation across these three aspects, this research aims to provide a comprehensive evaluation of the feasibility, reliability, and practical value of RAG in medical question answering, offering empirical evidence to support the development and optimization of medical AI systems.

3 Methodology

3.1 Dataset

The dataset used in this study is sourced from the Healthcare Dataset available on the Kaggle platform. This dataset is characterized by its broad coverage, containing a variety of medical question-answer pairs and knowledge entries spanning topics such as common diseases, symptom descriptions, and treatment methods. Its moderate size makes it well-suited for training and evaluating medical question-answering systems while maintaining data controllability and processing efficiency. Due to constraints in computational resources and time, this study did not employ the full dataset. Instead, the focus was placed specifically on cancer-related data. Cancer-related questions and answers carry particular clinical relevance and representativeness in the medical field, making them valuable for supporting related research. Limiting the scope of the data also helps ensure better experimental control and reproducibility under limited resource conditions. Therefore, the selection of this subset balances both the professional demands of medical Q&A and the practical requirements of conducting structured experiments within resource limitations.

3.2 Models and Methods

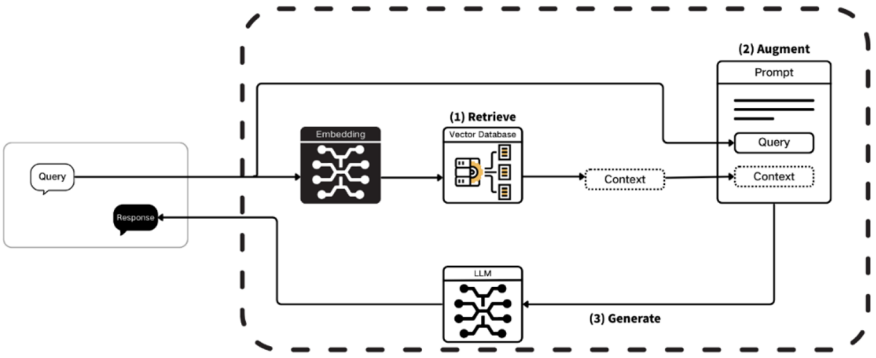

Fig. 1. Workflow of Retrieval-Augmented Generation (RAG) Model

The present study employs a RAG framework to construct a medical question-answering system, addressing issues such as hallucinations and lack of traceability that large language models often exhibit in medical contexts. As shown in Figure 1, the RAG process is comprised of three principal components: the construction of a knowledge base, the retrieval module, and the generation module.

3.2.1 Knowledge Base Construction

Initially, the selected cancer-related medical data underwent a series of preprocessing steps, encompassing tokenization, the elimination of noise, and the harmonisation of medical terminology. Subsequently, a pre-trained embedding model (for example, All-MiniLM-L12-v2) was employed to encode the medical question–answer pairs into high-dimensional vectors. The FAISS algorithm was utilised to construct an efficient vector index, thereby facilitating expeditious similarity retrieval over extensive knowledge entries [13]. The resulting knowledge base provides authoritative and structured medical information to support subsequent question-answering tasks.

3.2.2 Retrieval Module

Upon receiving a user query, the system first encodes the question into a vector and retrieves the top-k most relevant medical documents from the knowledge base. The retrieval process is facilitated by a cosine similarity–based nearest neighbour search method [14], thereby ensuring that the selected documents are semantically closest to the user's question. The present study places particular emphasis on the choice of the retrieval parameter *k*, conducting experiments under four configurations (k=1, 3, 5, and 8) to analyse the impact of different retrieval scales on model performance. The retrieval module's objective is to supply the generation module with high-quality candidate knowledge, thereby reducing the likelihood of the model fabricating answers.

3.2.3 Generation Module

Following the retrieval of the relevant documents, the system concatenates the user query with the retrieved results into a prompt, which is then fed into a large language model for answer generation. In this study, Qwen-Plus and Qwen-Flash are selected as the generation modules, both of which represent strong-performing large language models in the current open-source ecosystem. The RAG mechanism facilitates the joint utilisation of the retrieved semantic information and its inherent language modelling capabilities, thereby producing responses that are more accurate, domain-specific, and interpretable.

3.2.4 Baseline Comparison and Experimental Setup

In order to demonstrate the efficacy of the RAG framework, the experiments were designed with two types of comparison. The first involved baseline models, where Qwen-Plus and Qwen-Flash were directly applied to generate answers without the support of any external retrieval. The second employed the RAG-enhanced models, in which the retrieval–generation pipeline was utilized and answer generation was conducted under different top-k configurations.

In addition, to ensure fairness, all experiments were carried out under the same hardware environment, with model parameters kept at their default settings and no additional fine-

tuning applied. It is evident that the design in question is such that performance differences are primarily attributable to the RAG framework and retrieval parameters.

3.3 Experimental Design

In order to systematically evaluate the feasibility and reliability of RAG in the context of medical question answering tasks, this study conducts experiments from three perspectives: the mitigation of hallucinations, the attribution of sources, and the impact of retrieval parameters. The experimental design is comprised of three constituent components: task setup, evaluation metrics and experimental schemes.

3.3.1 Task Setup

The primary objective of the experiments is Medical Question Answering (MedQA). The input comprises medical questions in natural language, and the output is the model-generated answers. In order to ensure the study's specificity, all questions were derived from curated cancer-related medical data. Two categories of models were designed for the experiments: baseline models and RAG models.

To comprehensively assess model performance, this study adopts a set of multidimensional quantitative metrics: BERTScore F1: Measures the semantic similarity between the model-generated answers and the reference answers, reflecting accuracy and domain expertise. Refusal Rate: Calculates the proportion of instances where the model chooses to decline answering when faced with unanswerable questions or insufficient knowledge, serving as an indicator of hallucination suppression. Attribution Accuracy: Examines whether the generated responses are grounded in the retrieved medical documents, evaluating answer traceability and interpretability.

3.3.2 Experimental Protocol

In the experimental design, this paper first compares the baseline model with the RAG model through comparative experiments across three metrics: accuracy, source traceability, and refusal capability. This finding serves to substantiate the efficacy of RAG in the medical domain. Subsequently, experiments were conducted to ascertain the sensitivity of the model to varying parameters. This was achieved by setting different retrieval scales (top-k = 1, 3, 5, 8) within the RAG model. The impact of retrieval parameters on model performance was analysed in order to explore a reasonable retrieval window. Finally, multi-model experiments are conducted by running the same experimental protocol on both Qwen-Plus and Qwen-Flash to validate the robustness of the findings. This study employs a systematic design to comprehensively evaluate the feasibility and reliability of RAG in medical question-answering. This evaluation is conducted from multiple perspectives, thereby providing a reference for the optimisation of future medical intelligent systems.

4 Implementation

The experiments in this study were conducted on a computer equipped with an AMD Ryzen 9 7945HX processor (2.50 GHz), 16 GB of RAM, and an NVIDIA RTX 4060 GPU, running the Windows 11 operating system. The primary software environment comprised Python 3.10, PyTorch 2.0, HuggingFace Transformers, Sentence-Transformers, FAISS, and LangChain. Specifically, Sentence-Transformers was employed for text embedding, FAISS for constructing an efficient vector retrieval index, and LangChain for integrating retrieval

and generation modules to build a complete RAG framework. For the substantial language model component, the APIs of Qwen-Plus and Qwen-Flash were employed to ensure seamless collaboration between the generation and retrieval modules.

During the data processing and experimental implementation phases of this study, the present paper screened cancer-related entries from Kaggle medical datasets. Following a thorough cleaning and vectorisation process, a comprehensive medical knowledge base was constructed, and FAISS was employed to facilitate efficient retrieval. The system integrates user queries with retrieved results, subsequently utilising this data to generate answers for both Qwen-Plus and Qwen-Flash. The impact of retrieval scale on performance is analysed by different top-k parameters. Concurrent experimentation on baseline and RAG models is conducted, with rigorous evaluation through BERTScore, refusal rate, and source verification rate to ensure the reliability and reproducibility of research findings.

5 Results and Analysis

The present study set out to evaluate the effectiveness and reliability of RAG in the context of medical question-answering. To this end, experimental analyses were conducted across three dimensions: answer quality, impact of retrieval parameters, and system credibility.

Firstly, with regard to the quality of the responses, as demonstrated in Figure 2, the results indicate that RAG significantly enhances the semantic accuracy of the generated responses. Utilising BERTScore F1 as the evaluation metric, the baseline model (lacking external retrieval) attained scores of 0.668 and 0.653 on Qwen-Plus and Qwen-Flash, respectively. Following the incorporation of RAG, the scores demonstrated a marked improvement, reaching 0.701 and 0.672, respectively, signifying substantial enhancements. This outcome validates the effectiveness of RAG, demonstrating that integrating retrieval mechanisms in medical question-answering mitigates language models' knowledge hallucination issues. This approach yields responses that are more closely aligned with standard answers, while concomitantly enhancing overall professionalism and semantic consistency.

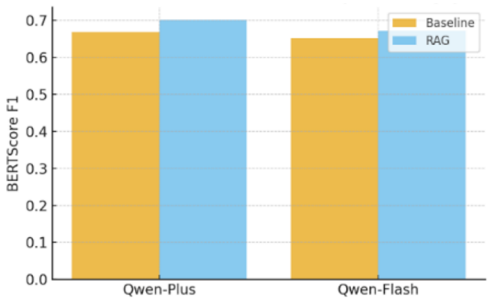

Fig. 2. Comparison of Baseline and RAG models on BERTScore F1 in medical question answering

Secondly, with regard to the impact of retrieval parameters on performance, experiments were conducted to compare outcomes under different top-k configurations. As demonstrated in Figure 3, the findings reveal that on Qwen-Plus, BERTScore initially rises and subsequently falls with increasing k values, attaining optimal performance at k=5 (0.710). A similar pattern emerges on Qwen-Flash, where the highest score (0.695) is also achieved at k=5. This indicates that excessively small k values may result in inadequate knowledge coverage, while excessively large k values may introduce redundant or noisy documents, thereby compromising the quality of the answers. Consequently, a reasonable retrieval scale is a critical factor influencing RAG performance. The findings of this study provide a reference for the "optimal retrieval window" in the design of medical question-answering systems.

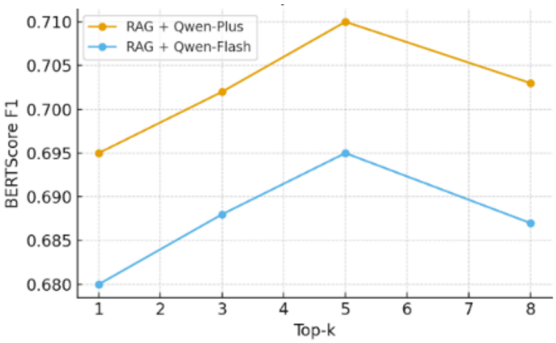

Fig. 3. Effect of different top-k values on BERTScore performance for RAG models

Finally, in terms of system reliability and credibility, RAG's performance in refusal rate and source verification rate further demonstrates its potential in medical applications. As demonstrated in Figure 4, the refusal experiments indicate that RAG+Qwen-Plus and RAG+Qwen-Flash achieved refusal rates of 77% and 71%, respectively. This suggests that the models prefer to refuse answering questions they cannot address or lack sufficient knowledge about rather than fabricating responses, thereby effectively mitigating hallucination risks. As demonstrated in Figure 5, with regard to source verification, RAG+Qwen-Plus attained a verification rate of 73%, while RAG+Qwen-Flash achieved 65%, suggesting that the majority of responses could be traced back to authentic retrieved documents. The findings of this study corroborate the hypothesis that RAG enhances not only the accuracy of responses in the context of medical question-answering but also the interpretability and credibility of the system. This provides substantial empirical evidence that lends credence to the prospective utilisation of medical AI systems in clinical settings.

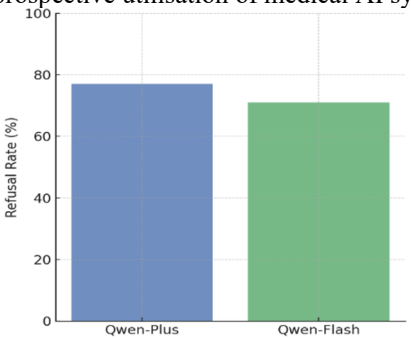

Fig. 4. Refusal rates of RAG models (Qwen-Plus and Qwen-Flash)

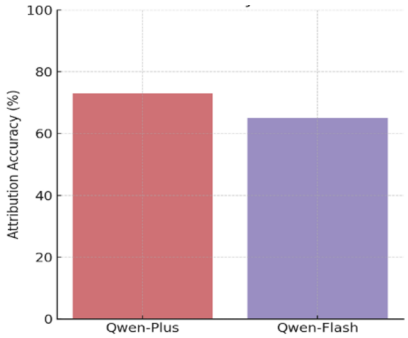

Fig. 5. Attribution accuracy of RAG models (Qwen-Plus and Qwen-Flash)

In summary, the experimental results provide validation of RAG's advantages in medical question-answering across three dimensions: it effectively enhances answer quality, reveals the significant impact of retrieval parameters on performance, and demonstrates favourable characteristics in terms of credibility and interpretability. The findings of the study provide practical guidance for the design of RAG systems in medical contexts and lay the groundwork for the exploration of trustworthy AI applications in high-stakes domains.

6 Discussion and Future Work

Although this study validates the effectiveness and reliability of RAG in medical question answering, several limitations remain. First, due to constraints in computational resources and time, experiments were conducted only on cancer-related questions. While this focus is representative, the conclusions may not generalize well to other disease areas. Secondly, all experiments were carried out using a single consumer-level GPU (NVIDIA RTX 4060), which has limited memory capacity. This imposed constraints on both the size of the constructed knowledge base and the complexity of the models that could be deployed. Thirdly, the study primarily employed the open-source models Qwen-Plus and Qwen-Flash, excluding models with alternative architectures. Therefore, further verification is required to confirm the conclusions' broader applicability. Finally, due to budget and time constraints, it was not possible to significantly expand the knowledge base or carry out cross-domain validation, which affected the generalisability of the results to some extent.

Future studies could advance along several paths. Firstly, future works can test RAG on larger and more diverse medical datasets from different medical areas such as cardiology or neurology to evaluate how generalize our results will be. Secondly, the knowledge database should be enriched by multimodal medical data—for example, CT image, MRI image, EHR with structure fields, and more complex knowledge graphs—to make the model better answer complicated medical questions. Thirdly, we might think about even more powerful retrieval method—for example, context-based search or multi-stage hybrid search—but still, it would be very good for further exploration in improving precision of retrieving the most helpful content. Moreover, another direction is generating more readable and clearer responses. Additionally, in the future, comprehensive testing with adversarial examples, fuzzy inputs, multilingual inputs, etc., needs to be taken into account so that we can ensure RAG's robustness in the real world, which is key for reliable implementation in clinical practice. Lastly, at the practical level, integrating RAG with clinical decision support systems could help assess its utility in areas like diagnostic assistance, patient management, and medical training.

7 Conclusion

This study investigates the application of RAG in medical question-answering systems, with the aim of examining its effectiveness and reliability in reducing hallucinations in large language models, improving knowledge traceability, and optimizing parameter configurations. A medical knowledge base was constructed using cancer-related questions from Kaggle medical datasets. The Qwen-Plus and Qwen-Flash models were employed as generation modules to compare and analyze the performance of RAG against baseline models. Experimental results demonstrate that RAG offers significant advantages in medical question-answering tasks. First, in terms of response quality, RAG models outperform baseline models on the BERTScore F1 metric, indicating that incorporating a retrieval mechanism effectively mitigates hallucination issues and generates responses that align more closely with reference answers. Second, regarding parameter tuning, different top-k

configurations considerably affect performance. The findings suggest that optimal performance is achieved when k=5, providing valuable insights for developing medical question-answering systems. Finally, in terms of system reliability, RAG models exhibit high refusal rates and source verifiability. When confronted with uncertain questions, they adopt a cautious approach by providing answers traceable to their sources, thereby enhancing the credibility and interpretability of medical question-answering systems. The study indicates that RAG shows strong feasibility and application potential in medical question-answering scenarios, offering empirical support for the development and optimization of medical AI systems. However, due to constraints in computational resources, time, and funding, the validation was conducted only on cancer-related questions. Future research is recommended to expand to larger scales, cover more disease areas, and incorporate more diverse large language models. Furthermore, integrating domain-specific knowledge graphs, multimodal medical data, and more advanced retrieval strategies represents a promising direction for further exploration.

References

1. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. 43, 2, Article 42 (March 2025), 55 pages. https://doi.org/10.1145/3703155
2. D. Roustan., and F. Bastardot. The clinicians' guide to large language models: A general perspective with a focus on hallucinations. Interact J Med Res 2025. 14, e59823 (2025). https://doi.org/10.2196/59823
3. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. (2020, May 22). https://arxiv.org/abs/2005.11401
4. Retrieval Augmented Generation (RAG) for LLMs – Nextra. (n.d.). Prompting Guide. Retrieved September 3, (2025), from https://www.promptingguide.ai/research/rag
5. R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. Bitterman. et al. Retrieval-augmented generation for generative artificial intelligence in health care. npj Health Syst. 2, 2 (2025). https://doi.org/10.1038/s44401-024-00004-1
6. T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198. (2023). https://doi.org/10.1371/journal.pdig.0000198
7. R. Vaishya, A. Misra, A. Vaish, and A. Turki. Hallucinations in ChatGPT: A serious concern. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 17(4), 102744. (2023). https://doi.org/10.1016/j.dsx.2023.102744
8. G. Izacard, and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. arXiv. (2020, July 2). https://arxiv.org/abs/2007.01282
9. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. REALM: Retrieval-augmented language model pre-training. arXiv. (2020, February 10). https://arxiv.org/abs/2002.08909

10. W. Zhang, and J. Zhang. Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics, 13(5), 856. (2025). https://doi.org/10.3390/math13050856
11. G. Xiong, Q. Jin, Z. Lu, and A. Zhang. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics: ACL 2024, 6233–6251. (2024). https://doi.org/10.18653/v1/2024.findings-acl.372
12. Y. Shi, S. Xu, T. Yang, Z. Liu, T. Liu, Q. Li, X. Li, and N. Liu. MKRAG: Medical knowledge retrieval augmented generation for medical question answering. arXiv. (2023, September 27). https://arxiv.org/abs/2309.16035
13. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The FAISS library. arXiv. (2024). https://arxiv.org/abs/2401.08281
14. R. Upadhyay, M. Viviani. Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. Discov Computing 28, 27 (2025). https://doi.org/10.1007/s10791-025-09505-5