Development of a Medical Question-Answering System Using Large Language Models with Retrieval Augmented Generation and Prompt Engineering

University of Turku
Department of Computing
Master of Science Thesis
Health Technology
June 2025
Hammad Ahmed Tahir

Supervisors:
Tuukka Panula
Mohammad Feli
Alabi Rasheed Omobolaji

UNIVERSITY OF TURKU
Department of Computing

Hammad Ahmed Tahir: Development of a Medical Question-Answering System Using Large Language Models with Retrieval Augmented Generation and Prompt Engineering

Master of Science Thesis, 57 p., 17 app. p.
Health Technology
June 2025

This thesis develops and evaluates a Medical Question-Answering System (MQAS) that combines Retrieval-Augmented Generation (RAG) with advanced prompt engineering techniques to provide accurate, explainable, and safe answers to medical questions. Healthcare professionals often struggle to quickly access reliable medical information, and existing AI systems frequently produce hallucinations or lack transparency in their reasoning. To address these challenges, this research implements a RAG pipeline using GPT-4o-mini, MedEmbed-small-v0.1 embeddings, and Pinecone vector storage, enhanced with DSPy-based prompt engineering.

The study compares three prompt engineering approaches—zero-shot Chain of Thought (CoT), few-shot CoT, and ensembled few-shot CoT—using the PubMedQA_instruction dataset. Evaluation combines quantitative RAGAS metrics with qualitative assessments from licensed physicians. Results demonstrate that few-shot CoT consistently outperforms other approaches, achieving superior scores in answer relevancy (0.9514), faithfulness (0.7317), and answer correctness (0.6243). Clinician evaluations further validate the system's effectiveness, with particularly high ratings for explainability (0.905) and clinical accuracy (0.89).

The findings suggest that structured prompting techniques, particularly few-shot CoT, can significantly enhance the performance of medical question-answering systems by improving both factual accuracy and reasoning transparency. This research contributes to the advancement of healthcare AI by demonstrating a GDPR-compliant approach to developing medical QA systems that earn clinician trust through explainable, accurate responses. While limitations exist, including small clinician sample size and API budget constraints, this work establishes a foundation for future research into cost-effective prompting strategies and real-world clinical implementations that could meaningfully improve healthcare information access.

Keywords: Medical LLMs, Medical Question Answering Systems, Retrieval Augmented Generation, PubMed, RAGAS

Contents

List of Figures

List of Tables

1 Introduction

In this rapidly changing age of technology, the vastness of medical knowledge is ever increasing. However this makes the need of access to this knowledge even more challenging. The objective of this thesis is to explore the development of a Medical Question Answering System (MQAS) by leveraging the power of Large Language Models (LLMs). It also highlights the role of prompt engineering in boosting accuracy, reliability and accessibility of such systems.

1.1 Background and Motivation

In the current digital era, having access to precise and reliable medical information is more crucial than ever for both patients and healthcare professionals. The rapid expansion of medical knowledge has led to over 30 million citations in PubMed and nearly 1 million new studies published each year [1]. This has created an overwhelming amount of information that makes it challenging for healthcare providers to keep up with the latest evidence-based practices. Likewise, patients trying to understand their health conditions encounter difficulties in navigating complex medical language and recognizing trustworthy sources amid the vast array of online health information.

Conventional methods of retrieving medical information, such as manually searching through databases like PubMed, MedlinePlus, and UpToDate, can be both time and resource consuming. These also often demand considerable domain knowledge

for effective interpretation [2]. Healthcare providers are estimated to spend approximately 5.5 hours each week looking for information, time that could be better spent on patient care [3]. Furthermore, the data sources for medical information are too scattered and diffused to be accessed directly in an efficient manner.

Recent progress in Artificial Intelligence (AI), especially concerning Large Language Models (LLMs), has revealed significant potential to change the way we access and understand medical information. Models like GPT-4 and Med-PaLM have shown remarkable abilities in comprehending natural language inquiries and producing responses that closely resemble those of humans when addressing intricate medical queries [2]. For example, Med-PaLM 2 achieved an accuracy rate of 86.5% on questions modeled after the United States Medical Licensing Examination (USMLE), showcasing a level of performance that is comparable to that of human medical professionals in specific scenarios.

With that being said, the application of LLMs in healthcare pose a variety of challenges. Although such models produce reasonable outputs that might seem factual but are always subject to hallucination which can cause serious consequences in the medical domain [4]. Furthermore, many models operate as "black boxes". The processes by which they produce outputs or whatever goes on in those "hidden layers" is largely unclear. Such complexity and lack of transparency lead to issues around explainability of these models which in turn penalise the trustworthiness of their applications in mission critical scenarios like in healthcare [5].

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address these challenges by grounding LLM responses in authoritative external sources [6]. By retrieving relevant information from trusted medical databases before generating responses, RAG systems can significantly reduce hallucinations and enhance factual accuracy. When combined with prompt engineering—the strategic design and optimization of inputs to guide LLM responses—RAG offers a power-

ful framework for developing reliable medical question-answering systems without requiring extensive model retraining [7].

Training large language models (LLMs) can be a resource intensive process, involving a huge amount of compute hours. Prompt Engineering techniques such as Chain of Thought (CoT), Few-Shot learning, and Self-Consistency methods, implemented within a RAG framework, have shown potential to improve LLM performance across various tasks. The integration of specialized medical embedding models like MedEmbed [8] further enhances the relevance and accuracy of retrieved information, creating a robust foundation for medical question answering.

1.2 Research Problem

Despite the promising capabilities of LLMs in medical question answering, existing Medical Question-Answering Systems (MQAS) face several critical challenges that limit their practical utility in healthcare settings:

1. **Accuracy and Reliability**: LLMs can generate plausible-sounding but factually incorrect information, a phenomenon known as hallucination. In the medical domain, where information accuracy directly impacts patient care, this poses significant risks [9].

2. **Lack of Explainability**: Many LLMs operate as "black boxes," making it difficult to understand how they arrive at specific conclusions or to verify the sources of their information. This lack of transparency undermines trust among healthcare professionals and patients [10].

3. **Ethical and Legal Concerns**: The use of AI in healthcare raises important questions about data privacy particularly regarding compliance with regulations like the General Data Protection Regulation (GDPR) and the Health

Insurance Portability and Accountability Act (HIPAA), informed consent, and algorithmic fairness [10].

4. **Integration with Existing Knowledge Sources**: Current systems often struggle to effectively integrate with authoritative medical databases and knowledge sources in real-time, limiting their ability to provide up-to-date and evidence-based information [7].

5. **Domain Adaptation**: General-purpose LLMs may not adequately capture the nuances and specialized terminology of medical language without extensive domain-specific training, which is resource-intensive and may not be feasible for all applications [2].

While fine-tuning LLMs on medical data has shown promise in addressing some of these challenges, this approach requires significant computational resources, large amounts of high-quality medical data, and must be repeated whenever the base model is updated. RAG combined with strategic prompt engineering offers an alternative approach that may enhance LLM performance in medical contexts without the need for extensive retraining [6]. However, research on optimal RAG implementations and prompting strategies for medical applications remains limited.

This thesis aims to address these challenges by developing and evaluating a Medical Question-Answering System that leverages RAG and prompt engineering techniques to enhance the accuracy, trustworthiness, and accessibility of medical information provided by LLMs.

1.3 Research Objectives

The primary objective of this thesis is to develop a Medical Question-Answering System (MQAS) that utilizes LLM with a Retrieval-Augmented Generation (RAG)

pipeline in different prompt settings to enhance the accuracy, explainability, and clinical relevance of medical information. Specifically, this research aims to:

1. Design and implement a comprehensive RAG framework for medical question answering that grounds LLM responses in authoritative medical sources, reducing hallucinations and improving factual accuracy.

2. Evaluate the effectiveness of different prompt engineering approaches (Zero-Shot, Few-Shot, Chain of Thought, and Self-Consistency/Ensembled) in improving the performance of GPT-4o-mini within a RAG pipeline.

3. Develop methods to enhance the explainability and trustworthiness of AI-generated medical information through appropriate prompt design and grounding of responses to the authoritative medical databases.

4. Implement a comprehensive evaluation framework combining RAGAS metrics (context precision, faithfulness, answer relevancy, etc) with expert validation by clinicians to assess system performance across multiple dimensions.

5. Create a user-friendly interface that makes accurate medical information accessible to healthcare professionals, while clearly communicating the limitations and appropriate use cases of the system.

1.4 Research Questions

To guide this research, the following questions have been formulated:

Primary Research Question:

How can large language models (LLMs) with Retrieval-Augmented Generation (RAG) be leveraged to develop an accurate, explainable, and clinically relevant Medical Question-Answering System (MQAS)?

Secondary Research Questions:

1. How do different prompt engineering techniques (e.g., Chain of Thought, Few-Shot, Self-Consistency) affect the accuracy of medical QA responses?

2. What techniques can be used to evaluate and improve the trustworthiness and explainability of LLM-generated medical answers?

3. How can grounding of AI generated responses to medical databases (e.g., PubMedQA) enhance the performance of an MQAS?

1.5 Thesis Structure

This thesis is organized into six chapters:

Chapter 1: Introduction provides background information on the challenges of medical information retrieval, introduces the potential of LLMs in healthcare, outlines the research problem, objectives, and questions that guide this study.

Chapter 2: Literature Review examines existing research on medical information retrieval systems, the evolution and applications of LLMs in healthcare, prompt engineering techniques, current approaches to medical question-answering systems, and ethical and legal considerations in AI-driven healthcare.

Chapter 3: Methodology describes the research design, dataset selection and preparation, prompt engineering experiments, system architecture, and evaluation framework used in this study.

Chapter 4: Results presents the findings from the prompt engineering experiments and expert clinical evaluation while adhering to legal and ethical compliance guidelines.

Chapter 5: Discussion interprets the results and findings, discusses the developed system, acknowledges limitations, discusses implications for healthcare practice, and suggests directions for future research.

Chapter 6: Conclusion summarizes the key findings, highlights the contributions to the field, and provides recommendations for implementing AI-powered medical question-answering systems.

The thesis concludes with a comprehensive list of references and appendices containing supplementary materials such as sample prompts, system architecture diagrams, evaluation metrics details, and code snippets.

Table 1.1: Overview of Thesis Chapters

Chapter	**Description**
1. Introduction	Background, problem, objectives, and structure
2. Literature Review	Existing research and theoretical framework
3. Methodology	Research design and system development
4. Results	Experimental findings and evaluation
5. Discussion	Analysis and implications
6. Conclusion	Summary and recommendations

1.6 Use of Artificial Intelligence in Thesis

This thesis used artificial intelligence (AI) tools for specific tasks, while all core intellectual work is my own. ChatGPT refined text and improved sentence cohesion, enhancing clarity in the written drafts. Grok's DeepSearch and Consensus AI supported literature searches, identifying relevant papers on medical question-answering systems, retrieval-augmented generation, and prompt engineering for Literature Review 2. DeepSeek filtered PDF articles, extracting content related to methodology 3 and results 4. Beyond these tasks—text refinement, literature search, and content filtering—all research design, experiments, analysis, and conclusions are my original work, ensuring academic integrity and transparency at the University of Turku.

2 Literature Review

This chapter provides a comprehensive review of the literature relevant to the development of a Medical Question-Answering System (MQAS) using Retrieval-Augmented Generation (RAG) and prompt engineering techniques. The review begins with an overview of AI applications in healthcare, followed by an examination of medical question-answering systems, RAG methodologies, prompt engineering techniques, medical QA datasets, evaluation frameworks, and concludes by identifying research gaps that this thesis aims to address.

2.1 Introduction to AI in Healthcare

Artificial Intelligence (AI) has emerged as a transformative force in healthcare. It has revolutionized various aspects of medical practice from diagnostics to treatment planning and patient care. The integration of AI technologies into healthcare systems has shown significant potential in improving patient care, enhancing administrative processes and creating a better patient experience by reducing costs and clinical risks [11].

2.1.1 AI-Driven Clinical Decision Support

AI-driven clinical decision support systems have demonstrated remarkable capabilities in augmenting healthcare professionals' diagnostic and treatment decisions. These systems can analyze complex medical data, including electronic health records

(EHRs), medical imaging, and laboratory results, to provide evidence-based recommendations that enhance clinical decision-making [12]. For instance, AI algorithms have achieved diagnostic accuracy comparable to experienced clinicians in specific domains such as dermatology, ophthalmology, and radiology.

The World Economic Forum [13] identified six key ways AI is transforming healthcare: enhancing diagnostic accuracy, personalizing treatment plans, streamlining administrative workflows, accelerating drug discovery, improving patient engagement, and expanding healthcare access in underserved regions. These applications collectively contribute to a more efficient, effective, and equitable healthcare system.

It has been emphasized in a research that AI integration in clinical practice has led to significant improvements in early disease detection, treatment optimization, and predictive analytics for population health management [14]. It is highlighted that AI systems can identify subtle patterns in patient data that might be overlooked by human clinicians, potentially leading to earlier interventions and improved patient outcomes.

2.1.2 Challenges in Healthcare AI

Despite its promising potential, AI implementation in healthcare faces several significant challenges. Data interoperability remains a primary obstacle, as healthcare information is often scattered across different systems and formats, making it difficult to create comprehensive datasets for AI training [12]. Additionally, ensuring compliance with regulatory frameworks such as GDPR in Europe and HIPAA in the United States presents complex legal and ethical considerations for AI developers and healthcare organizations.

Trust issues also pose substantial barriers to AI adoption in clinical settings. [14] noted that healthcare professionals often express concerns about the abstract nature of many AI systems, where the reasoning behind recommendations is not transparent

or explainable. This also may vary among people depending upon socioeconomic and demographic parameters. This lack of explainability can undermine clinician confidence and hinder the integration of AI tools into routine clinical practice.

Integrating LLMs into existing clinical systems is complex, especially given the diverse software environments in healthcare [15]. Regulatory issues also play a major role. In many jurisdictions, LLM-powered tools may be subject to medical device regulations. The opaque decision-making of these models further complicates approval, as agencies typically require transparency and demonstrable safety [10].

2.2 Medical Question-Answering Systems

Medical Question-Answering Systems (MQAS) represent a specialized application of AI in healthcare, designed to provide accurate and relevant answers to medical queries posed by healthcare professionals and patients. These systems have evolved significantly over time, from rule-based approaches to sophisticated machine learning models and, most recently, large language models (LLMs).

2.2.1 Early Medical QA Systems

Early medical QA systems primarily relied on rule-based approaches and structured knowledge bases. These systems utilized predefined rules and medical ontologies to match questions with appropriate answers from curated knowledge sources [16]. While these systems provided reasonable accuracy for well-defined question types, they lacked flexibility in handling natural language variations and struggled with complex or ambiguous queries.

As machine learning techniques advanced, statistical approaches began to supplement rule-based systems, enabling more sophisticated natural language understanding and improved answer retrieval. However, these early ML-based systems

still required extensive feature engineering and domain-specific customization, limiting their scalability and adaptability to new medical domains [17].

2.2.2 Modern LLM-Based QA Systems

Deep learning has brought a notable advancement in Medical Question Answering. Models such as BioBERT [18] and ClinicalBERT [19], fine-tuned with biomedical and clinical texts, excelled in tasks like named entity recognition and question answering. These extractive systems identified text segments that addressed specific queries but faced challenges with intricate inquiries that required information synthesis across various documents or inference.

LLMs are making headlines in achieving human performance in analytical and reasoning tasks. Models such as Med-PaLM 2 have demonstrated remarkable capabilities in understanding and responding to complex medical queries. [2] reported that Med-PaLM 2 achieved performance comparable to human medical experts i.e. about 86.5% accuracy, including USMLE-style questions, representing a significant advancement in the field.

Present-day LLM-based MQAS exhibit variations in several respects. Some general purpose models like GPT-4 through their remarkable baseline performance show immense potential in medical domain. Innovations in prompt engineering and dynamic retrieval systems, further distinguish the performance of these systems. BioGPT [20] is also a significant advancement in biomedical NLP, combining generative capabilities with domain-specific pre-training to achieve State of the Art (SOTA) results. Its success highlights the importance of in-domain training and tailored architectures for specialized applications.

Another crucial development is the merging of LLMs with structured medical knowledge. GatorTron [21], for instance, integrates LLMs with medical ontologies like UMLS to deepen the understanding of medical concepts and their interrelations.

This strategy aids in improving both precision and interpretability, addressing the shortcomings in the conceptual grounding of LLMs.

Multimodal capabilities signify a new area of advancement. Models like GPT-4V can handle both text and images, facilitating answers to visual medical questions, interpreting diagnostic tests, and recognizing symptoms from photographs. Although these tools are still in development, they have the potential to significantly expand the applications of MQAS.

2.2.3 Gaps in Medical QA

Several significant gaps remain in current medical QA systems. [16] identified explainability as a critical limitation, noting that many systems provide answers without clear reasoning or evidence trails, making it difficult for users to assess the reliability of the information. This lack of transparency is particularly problematic in healthcare, where understanding the rationale behind recommendations is essential for clinical decision-making. Future research in this area should focus on Explainable AI Techniques (XAI).

Real-world validation represents another significant gap. [2] emphasized that while many systems perform well on benchmark datasets, however their performance in actual clinical settings with diverse patient populations and complex, ambiguous queries remains inadequately evaluated. This gap highlights the need for more comprehensive evaluation frameworks that incorporate feedback from healthcare professionals and assess performance across diverse clinical scenarios.

Additionally, [17] noted that many existing systems struggle with context-awareness and personalization, providing generic answers rather than responses tailored to specific patient characteristics or clinical contexts. Addressing these gaps requires innovative approaches that combine the flexibility of LLMs with robust knowledge retrieval and context-aware reasoning.

2.3 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address the limitations of standalone LLMs, particularly in knowledge-intensive domains like healthcare. By combining the generative capabilities of LLMs with explicit retrieval from authoritative external sources, RAG systems can provide more accurate, up-to-date, and verifiable responses to medical queries.

2.3.1 Principles of RAG

The fundamental principle of RAG involves a two-stage process: retrieval followed by generation. As described in [6], the retrieval component identifies relevant documents or passages from a knowledge base in response to a query, while the generation component synthesizes this retrieved information along with the query to produce a coherent and accurate response.

This approach offers several advantages over standalone LLMs. First, it grounds responses in external knowledge sources, reducing hallucinations and factual errors. Second, it enables access to specialized or up-to-date information that may not be present in the LLM's training data. Third, it provides a degree of explainability by linking generated responses to specific source documents, enhancing transparency and trustworthiness.

2.3.2 RAG in Biomedical Applications

RAG has shown particular promise in biomedical applications, where accuracy and evidence-based responses are paramount. The integration of RAG with biomedical knowledge sources like PubMed has enabled more precise and reliable medical question answering. [6] demonstrated that RAG models outperformed both retrieval-only and generation-only baselines on knowledge-intensive tasks, including

biomedical question answering.

The effectiveness of RAG in biomedical contexts depends significantly on the quality and relevance of the retrieved information. Specialized embedding models for medical text, such as MedEmbed, have been developed to enhance retrieval performance by better capturing the semantic relationships between medical queries and documents [8]. These domain-specific embeddings help ensure that the most relevant medical information is retrieved and incorporated into the generated responses.

2.3.3 Recent Advances in RAG

Recent advances in RAG have further enhanced its capabilities for medical applications. Google Research [22] highlighted the importance of "sufficient context" in RAG systems, emphasizing that the quality and comprehensiveness of retrieved information significantly impact response accuracy. Their research demonstrated that adaptive retrieval strategies, which dynamically adjust the amount and type of information retrieved based on query complexity, can substantially improve performance on challenging medical questions.

Hybrid search approaches, combining dense vector retrieval with traditional keyword-based methods, have also emerged as a promising direction. [23] reported that hybrid search can capture both semantic relationships and specific medical terminology, leading to more comprehensive and relevant retrieval results. Additionally, real-time retrieval capabilities have been developed to incorporate the latest medical research and guidelines into RAG systems, addressing the challenge of information currency in rapidly evolving medical fields.

Another significant advancement is the development of multi-hop RAG architectures, which can follow chains of reasoning across multiple documents to answer complex medical questions that require synthesizing information from diverse

sources. These architectures are particularly valuable for addressing questions that involve multiple medical concepts or require integrating information across different specialties [22].

2.4 Prompt Engineering in AI Systems

Prompt engineering has emerged as a critical discipline for optimizing the performance of LLMs across various applications, including medical question answering. By strategically designing the inputs provided to LLMs, prompt engineering techniques can significantly enhance response quality, accuracy, and relevance without requiring model retraining or fine-tuning.

2.4.1 Basics of Prompt Engineering

At its core, prompt engineering involves crafting effective instructions, examples, and context to guide LLM behavior. [24] demonstrated that LLMs possess remarkable in-context learning capabilities, allowing them to adapt to specific tasks based solely on the prompts they receive. This discovery highlighted the importance of prompt design as a key factor in LLM performance. e.g. in this study GPT-3 could perform tasks like translation, question answering, and classification using only natural language prompts

Effective prompts typically include clear instructions, relevant context, and appropriate formatting to elicit the desired response [25]. For medical applications, prompts often incorporate domain-specific terminology, contextual information about the medical query, and explicit guidance regarding the expected response format and level of detail.

2.4.2 Zero-Shot and Few-Shot Learning

Zero-shot learning enables LLMs to complete tasks based solely on instructions, without examples. This technique is quite useful in assessing the baseline performance of LLMs. In a healthcare setting, a prompt like "Explain the symptoms of myocardial infarction" illustrates how zero-shot learning draws on pre-trained knowledge to generate informative responses. This approach is especially useful in healthcare, where creating labeled datasets is difficult due to privacy constraints and annotation costs. Its flexibility makes it easier to adapt to diverse tasks while avoiding biases that might arise from poorly chosen examples [26].

Few-shot learning enhances prompt engineering by including a handful of examples (usually 2–5) to demonstrate task structure and expected output [24]. For instance, a prompt might include several question-answer pairs before introducing a new case. This method helps the model infer patterns and produce more accurate, context-appropriate responses through implicit meta-learning.

Few-shot performance generally improves with model scale, making it increasingly viable as LLMs grow in size [24]. In medicine, providing a few curated examples has improved diagnostic performance [27]. Still, example selection is critical. [28] found that diversity, relevance, and clarity of examples strongly affect performance, and that curation often requires domain knowledge.

The order of examples also matters. [29] observed recency effects, where examples nearer to the query influence the response more. This introduces further complexity in designing prompts within limited context windows. Moreover, the consistency and structure of examples—such as clearly separating inputs from outputs—improve task comprehension.

Comparative studies provide insight into both approaches. [30] assessed LLM consistency in medical contexts using different prompting strategies. They found wide variation in agreement with medical guidelines (Fleiss kappa from -0.002 to

0.984), with few-shot prompts often enhancing reliability. However, the optimal method varied by model and task, emphasizing the need for tailored strategies.

Zero-shot prompts offer simplicity and flexibility but may fall short for complex tasks. Few-shot prompts often deliver better accuracy but at the cost of careful example design and context space. Hybrid approaches, like "zero-shot chain of thought" [31] attempt to combine the strengths of multiple techniques, offering promising directions for robust prompt engineering in healthcare.

2.4.3 Advanced Techniques

Several advanced prompt engineering techniques have been developed to enhance LLM performance on complex tasks. Chain of Thought (CoT) prompting [28], encourages step-by-step reasoning by explicitly instructing the model to "think step by step through" a problem before providing a final answer. This approach has proven particularly effective for medical questions that require multi-step reasoning or the application of clinical guidelines.

In healthcare, CoT prompting has shown strong potential. A review identified CoT as the most frequently used prompt design in medical applications, cited in 17 of 114 studies [7] . CoT prompting can be applied using either few-shot or zero-shot approaches. In few-shot CoT, the prompt includes examples showing how a problem is solved step by step. [32] found CoT improved GPT-4's performance on USMLE-style questions, especially in multi-step cases. [27] integrated CoT into Med-PaLM 2, contributing to its high scores on medical benchmarks. These findings suggest CoT helps bridge the gap between general LLM abilities and the specialized reasoning needed for clinical decision-making.

Self-consistency methods represent another significant advancement. It involves generating multiple outputs for the same prompt and selecting the most consistent answer, typically through majority voting or aggregation [33]. This technique

has been shown to improve accuracy on complex reasoning tasks by leveraging the model's ability to approach problems from different angles. In healthcare, self-consistency is found to be valuable in clinical decision-making. [30] assessed the reliability of LLMs under different prompting methods and observed that combining prompt design with self-consistency improved answer stability and consistency in medical contexts.

Another similar prompting technique is "Tree of Thoughts" [34]. It explores different reasoning branches in parallel, allowing the model to revise its approach if one path fails. "Least-to-most prompting" [35] decomposes problems into sequential sub-tasks, building up toward a complete solution.

Despite its strengths, self-consistency has limitations. Generating multiple responses increases computational costs, potentially limiting real-time applications. If the model has systemic biases, these can persist across generations, leading to consistently wrong outputs. Majority voting may also obscure valuable minority perspectives, especially on nuanced issues.

2.4.4 Applications in Healthcare

In healthcare contexts, prompt engineering techniques have been adapted to address the specific challenges of medical question answering. For diagnostic reasoning tasks, CoT prompting has been particularly valuable, as it mirrors the systematic differential diagnosis process used by clinicians. By encouraging LLMs to consider symptoms, risk factors, and potential diagnoses in a structured manner, CoT prompting can lead to more thorough and clinically sound responses [28].

Safety-oriented prompting represents another important application in healthcare, with prompts explicitly instructing models to acknowledge limitations, avoid making definitive claims outside their expertise, and emphasize the importance of professional medical advice. These approaches help mitigate potential risks asso-

ciated with AI-generated medical information while maintaining the utility of the system as an informational resource.

MED-Prompt [36] is a novel prompt engineering framework tailored for healthcare applications, integrating structured prompts, context-aware instructions, and iterative refinement to enhance clinical accuracy and relevance. Their framework, tested on datasets like PubMedQA, employs zero-shot, few-shot, chain-of-thought, and role-based prompts (e.g., "act as a cardiologist"), achieving a higher accuracy compared to baseline methods. Clinician feedback validated the framework's explainability and safety, critical for clinical trust. This approach directly informs the DSPy-based prompting strategy in this thesis, which leverages similar prompt types to improve the MQAS's performance on PubMedQA_instruction, ensuring accurate, explainable, and safe responses for clinicians and patients.

2.5 Medical QA Datasets

The development and evaluation of Medical Question-Answering Systems rely heavily on high-quality datasets that capture the complexity and diversity of medical queries. These datasets serve as training resources, benchmarks for performance evaluation, and sources of knowledge for retrieval-based systems.

2.5.1 Overview of Medical QA Datasets

Several specialized datasets have been developed for medical question answering, each with distinct characteristics and focus areas. [37] introduced PubMedQA, a dataset derived from PubMed abstracts that includes research-oriented biomedical questions paired with evidence-based answers. MedQuAD, another prominent dataset, comprises question-answer pairs collected from trusted medical resources such as the National Library of Medicine and the Centers for Disease Control and

Prevention [38].

BioASQ is another notable dataset containing medical questions and ground truth answers along with related materials. This is a manually curated dataset that focuses on real-world information needs of the clinical experts. Unlike conventional question answer datasets which contain exact answers, this dataset relies on combining both structured and unstructured data. It further assists knowledge retrieval through the use of linked documents and snippets with each question-answer pair [39].

2.5.2 PubMedQA Instruction Dataset

The PubMedQA_instruction dataset, an instruction-tuned version of the original PubMedQA, has emerged as a particularly valuable resource for developing RAG-based medical QA systems. [37] described the dataset's structure, which includes clear question-answer pairs derived from PubMed abstracts, making it well-suited for both training and evaluation purposes.

What makes PubMedQA_instruction especially suitable for RAG applications is its direct connection to the biomedical literature. Each question-answer pair is linked to specific PubMed abstracts, providing a natural foundation for retrieval-based approaches that ground responses in authoritative medical sources. This structure enables the development of systems that can not only generate accurate answers but also provide evidence and citations to support their responses.

2.5.3 Challenges and Gaps

Recent advances in medical QA datasets have driven progress, but several persistent challenges still limit their effectiveness in building robust systems. A key issue is domain specificity—most datasets focus narrowly on certain medical specialties or question types, reducing their applicability across broader healthcare contexts. [40]

emphasized the need for more diverse datasets that span multiple disciplines and question formats.

Bias and equity concerns are also prominent. [41] systematically evaluated biases in retrieval-augmented medical QA systems, finding that demographic-sensitive queries often resulted in retrieval discrepancies that propagated through to final answers.

Another major limitation is annotation quality. [2] noted that many datasets rely on non-expert or automated annotations, which lack clinical expertise. Their work with Med-PaLM 2 showed that smaller, expert-validated datasets produced more accurate and clinically relevant responses than larger, less reliable ones.

The temporal nature of medical knowledge also poses a challenge. [16] found many datasets quickly become outdated, stressing the need for temporal markers and regular updates to reflect evolving medical evidence.

Finally, resource constraints hinder dataset creation. [42] highlighted the difficulty of obtaining expert annotations and normalizing medical concepts. They suggested collaborative frameworks and semi-supervised methods to ease the burden, though large-scale dataset development remains costly.

To tackle these issues, researchers have proposed subset selection with quality control [2] and hybrid approaches that blend synthetic data with expert validation [16]. These innovations aim to balance scale, diversity, and clinical accuracy in future medical QA datasets.

2.6 Evaluation Frameworks for Medical AI

Robust evaluation frameworks are essential for assessing the performance, reliability, and clinical utility of medical AI systems, including question-answering applications. These frameworks typically combine quantitative metrics with qualitative assessments to provide a comprehensive understanding of system capabilities and

limitations.

2.6.1 Quantitative Evaluation Metrics

Several quantitative metrics have been developed to evaluate different aspects of medical QA systems. For multiple-choice medical questions or yes/no questions, accuracy or F1 scoring is the most commonly used metric. This has been central in evaluating models like Med-PaLM 2 on datasets such as MedQA and MedMCQA [43]. Evaluating open-ended medical questions is complex due to their variability and the need for clinical precision. Metrics like BLEU, ROUGE, and METEOR, as noted by [38], measure textual overlap with reference answers, focusing on word or phrase similarity. However, these tools often fail to assess clinical accuracy or relevance, prioritizing surface-level form over medical substance, which limits their effectiveness in high-stakes medical contexts.

To address these shortcomings, domain-specific methods have emerged. [38] introduced a question-entailment approach, leveraging semantic resources to align responses with medical concepts, even if phrased differently. Similarly, [27] employed factual consistency checks, comparing outputs to trusted sources to identify hallucinations or misinformation, enhancing the reliability of medical question-answering systems.

The Retrieval-Augmented Generation Assessment (RAGAS) framework, introduced by [44], provides a comprehensive set of metrics specifically designed for evaluating RAG systems. These metrics include context precision (relevance of retrieved information), faithfulness (factual consistency between retrieved context and generated answer), answer relevancy (alignment with the query), and context recall (coverage of necessary information).

2.6.2 Qualitative Evaluation Methods

Automated metrics alone aren't enough. Human evaluation plays a crucial role. [43] designed a framework to assess responses across dimensions like factual accuracy, reasoning, comprehensiveness, potential harm, and bias. Their evaluations included both physicians and non-experts to capture varied perspectives. This approach highlights that effective responses must be not only correct but also safe, clear, and audience-appropriate.

Clinical relevance is another key concern. A model might generate an answer that appears accurate but omits critical clinical details or misprioritizes information. Expert reviewers are often needed to assess the practical usefulness of responses in real medical contexts. Though resource-intensive, this provides essential insight into system utility in healthcare settings.

Safety is particularly critical in MQAS. Inaccurate responses could lead to harmful outcomes. To address this, [27] evaluated responses for risks of physical, emotional, or financial harm. Identifying and addressing unsafe content is vital before deployment in real-world environments.

Explainability is also important for trust. As noted by [7], systems that justify their answers, cite sources, and acknowledge uncertainty are more trusted by healthcare professionals. Evaluation should consider whether responses include sound reasoning and reliable references.

2.6.3 Recent Evaluation Frameworks

Recent years have seen the development of more comprehensive evaluation frameworks specifically designed for medical AI applications. The FDA outlined an evaluation framework for AI-enabled medical devices that emphasizes performance assessment across diverse patient populations, continuous monitoring, and transparency in reporting limitations and potential biases [45].

The Transparent Evaluation of Health AI (TEHAI) [46] framework represents another significant advancement, providing a structured approach to assessing AI systems across multiple dimensions, including technical performance, clinical validation, ethical considerations, and implementation feasibility. This multidimensional approach recognizes that medical AI evaluation must extend beyond technical metrics to consider the broader context of healthcare delivery and patient outcomes.

[44] highlighted the importance of automated evaluation tools that can efficiently assess large numbers of system outputs while maintaining alignment with human judgments. The RAGAS framework demonstrates how automated metrics can be designed to capture the specific requirements of medical applications, providing developers with actionable insights for system improvement.

2.7 Research Gaps and Thesis Contribution

The literature review has identified several significant gaps in current approaches to medical question answering, which this thesis aims to address through the development and evaluation of a novel Medical Question-Answering System using RAG and prompt engineering techniques.

2.7.1 Identified Gaps

A critical gap in the current literature is the limited exploration of optimized RAG implementations specifically tailored for medical question answering. While RAG has shown promise in knowledge-intensive domains, the specific requirements and challenges of medical applications—such as handling specialized terminology, integrating diverse knowledge sources, and ensuring clinical relevance—remain inadequately addressed in existing research.

Another significant gap is the scarcity of clinician-validated evaluations of med-

ical QA systems. [38] noted that many studies rely primarily on automated metrics or benchmark datasets, with limited assessment of how systems perform in realistic clinical scenarios or how their outputs are perceived by healthcare professionals. This gap highlights the need for evaluation approaches that combine technical metrics with structured feedback from medical experts.

Prompt engineering shows great promise in medical applications, particularly in Question Answering (QA). A review noted that 78/114 medical prompt engineering studies include Prompt Design (PD) i.e. manually crafting prompts to better guide LLMs [7]. Further research is required in examining explainability and factual accuracy of AI generated responses while addressing challenges in privacy compliance.

Retrieval Augmented Generation (RAG) improves medical question answering by mitigating the computational requirement of fine-tuning or retraining LLMs. MKRAG [47] injects relevant medical facts into the prompts to improve the accuracy of open source LLMs like Vicuna-7B from 44.46% to 48.54% on MedQA-USMILE dataset. Despite showing success, there is plenty of room in making the current static retrieval systems more dynamic and flexible to accommodate multi-domain queries in medicine.

2.7.2 Contribution of the MQAS

This thesis contributes to addressing these gaps through the development and evaluation of a Medical Question-Answering System that integrates several innovative approaches. The system leverages GPT-4o-mini within a RAG pipeline, combining the flexibility and natural language understanding capabilities of a state-of-the-art LLM with the factual grounding provided by retrieval from authoritative medical sources.

The implementation of DSPy [48] for automated prompt design and experimentation represents another significant contribution, enabling systematic comparison

of different prompt engineering techniques (Zero-Shot, Few-Shot, Chain of Thought, and Self-Consistency) within the context of medical question answering. This approach provides insights into how prompt design affects various aspects of system performance, from factual accuracy to explainability and clinical relevance.

Furthermore, the comprehensive evaluation framework developed for this thesis addresses the gap in clinician-validated assessments by combining quantitative metrics from the RAGAS framework with structured feedback from healthcare professionals. This mixed-methods approach provides a more nuanced understanding of system performance across multiple dimensions relevant to medical applications.

By addressing these gaps, this thesis contributes to the advancement of medical question-answering systems that can provide accurate, explainable, and clinically relevant information to healthcare professionals and patients, ultimately supporting improved healthcare decision-making and information access.

3 Methodology

This section outlines the research design, dataset preparation, system architecture, prompt engineering experiments, evaluation framework, and implementation details for developing an MQAS using GPT-4o-mini with a Retrieval-Augmented Generation (RAG) pipeline. The methodology addresses challenges in medical information retrieval, ensuring accuracy, explainability, and clinical relevance while complying with GDPR/HIPAA standards [10].

3.1 System Architecture

The MQAS is built on a Retrieval-Augmented Generation (RAG) pipeline implemented using Haystack, a framework designed for building production-ready search and question-answering systems. The architecture integrates several key components to create a robust and efficient medical question-answering system.

3.1.1 Pipeline Overview

The RAG pipeline utilizes GPT-4o-mini as the generator component, responsible for producing the final answers to medical queries. For document embedding and retrieval, the system employs MedEmbed-small-v0.1, a specialized embedding model for medical text [8]. These embeddings are stored in Pinecone, a vector database optimized for similarity search operations. Figure 3.1 shows the architecture of MQAS with retrieval and generation components.

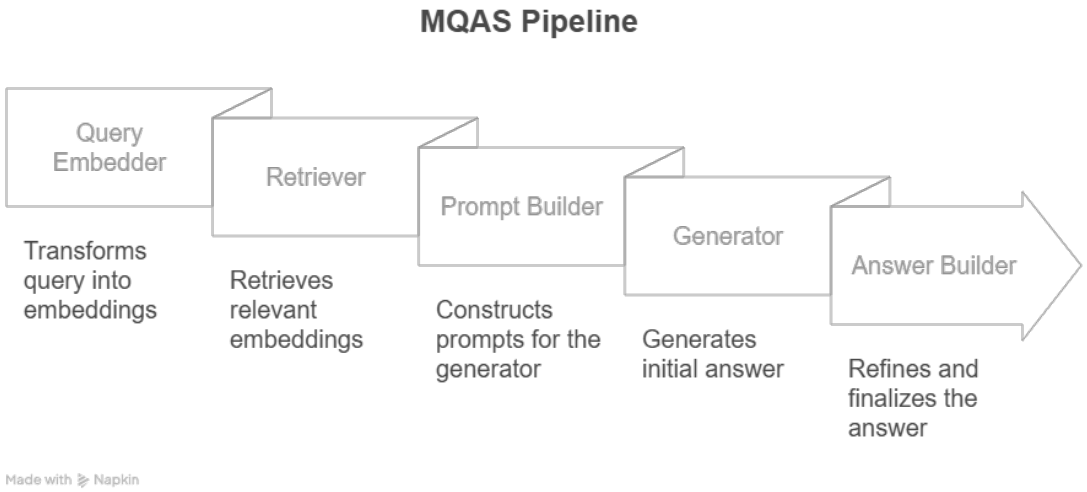

Figure 3.1: MQAS architrecture-integration of the system components.

As [6] demonstrated, this RAG approach significantly enhances performance in knowledge-intensive tasks by combining the generative capabilities of LLMs with explicit retrieval from external knowledge sources. The integration of these components creates a system that can generate accurate, contextually relevant responses to medical questions while maintaining factual grounding.

3.1.2 Component Integration

The system architecture consists of several integrated components that work together to process medical queries and generate accurate responses:

1. **EmbeddingRetriever**: This component converts incoming queries into vector embeddings using MedEmbed-small-v0.1 and retrieves relevant documents from the Pinecone vector store using similarity search.

2. **Generator**: GPT-4o-mini serves as the generation component, taking the retrieved documents and the original query as input to produce a comprehensive and accurate response.

The RAG process follows a sequential flow: the user query is first embedded using MedEmbed-small-v0.1, then relevant documents are retrieved from Pinecone based on embedding similarity, and finally, GPT-4o-mini generates a response that incorporates both the query context and the retrieved information. Table 3.1 lists all the components used in the MQAS along with their rationale.

Table 3.1: System Components and Their Functions

Component	**Technology**	**Function**
Generator	GPT-4o-mini	Answer generation
Embeddings	MedEmbed-small-v0.1	Query and document embedding
Vector Storage	Pinecone	Similarity search and retrieval
Framework	Haystack	Pipeline orchestration
Deployment	Docker-compose	System containerization
Frontend	OpenWebUI	User interface
API	FastAPI with Hayhooks	Backend service

3.1.3 Application Setup

The MQAS is served via Hayhooks with FastAPI, enabling OpenAI streaming for real-time responses. This implementation allows for efficient processing of medical queries with minimal latency, enhancing the user experience. The frontend utilizes OpenWebUI to provide an intuitive interface for both clinicians and patients, with appropriate adaptations for each user group.

The entire system is deployed using docker-compose, ensuring scalability, reproducibility, and ease of deployment across different environments. This containerized approach simplifies the management of dependencies and enables consistent performance across development and production environments. Appendix B shows the implementation of the RAG pipeline with all the components.

3.2 Dataset Selection and Preparation

The selection of an appropriate dataset is crucial for developing and evaluating a robust MQAS. This research utilizes the PubMedQA_instruction dataset from Hugging Face, which contains 272,000 training samples and 1,000 test samples. This dataset is selected for its comprehensive coverage of biomedical questions and its suitability for RAG implementation.

As [37] describe, the PubMedQA dataset focuses on research-oriented medical questions derived from PubMed abstracts, making it ideal for evaluating how well LLMs comprehend and synthesize information from biomedical literature. The instruction-tuned version of this dataset provides clear question-answer pairs that facilitate prompt engineering experiments and evaluation.

The subset of the training sample is created using DSPy's Example module, which facilitates the selection of a representative sample across different medical domains and question types. The subset selection process ensures that the experimental results will remain generalizable despite the reduced dataset size.

For efficient retrieval during the RAG process, document embeddings are stored in Pinecone, a vector database optimized for similarity search. This approach enables fast and accurate retrieval of relevant medical information during query processing. The Pinecone implementation is configured with Transport Layer Security (TLS) support to ensure security compliance, as recommended by [10] for handling sensitive medical information.

No additional preprocessing is required beyond the subset selection process, as the PubMedQA_instruction dataset is already well-structured for the experimental requirements. The dataset's existing format aligns well with the input requirements of the selected LLM and RAG pipeline.

The dataset preparation process ensures that the experimental foundation is robust, representative, and compliant with relevant regulations. The combination

Table 3.2: Dataset Characteristics

Characteristic	**Value**
Dataset Name	`PubMedQA_instruction`
Total Training Samples	272,000
Total Test Samples	1,000
Subset Size (Training)	1,000
Vector Storage	Pinecone
Embedding Model	MedEmbed-small-v0.1
Source	Hugging Face

of a biomedical-focused dataset with secure vector storage provides an appropriate basis for evaluating the performance of different prompt engineering techniques in a medical question-answering context.

3.3 Experimental Design

The research adopts an experimental approach with a clear objective: to develop an MQAS using GPT-4o-mini with RAG to enhance medical query accuracy, explainability, and clinical relevance. This objective directly addresses the research questions regarding the effectiveness of different prompt engineering techniques and the clinical reliability of LLM-generated medical responses.

The rationale for selecting GPT-4o-mini stems from its general-purpose capabilities which, when augmented by RAG, provide the necessary flexibility and accuracy for medical question-answering tasks [49]. While specialized medical models exist, GPT-4o-mini offers a cost-effective balance between performance and resource requirements, making it suitable for this research context. As [6] demonstrated, RAG significantly enhances performance in knowledge-intensive tasks by grounding LLM responses in authoritative external sources, addressing the critical issue of hallucination in medical contexts.

The experimental process follows an iterative methodology consisting of several sequential phases. First, the PubMedQA_instruction dataset is selected for its

biomedical focus and suitability for RAG implementation. Due to computational constraints, a subset of 1,000 training samples is carefully selected to ensure representative coverage while remaining computationally feasible.

The core experimental component involved systematically testing various prompt engineering techniques to optimize LLM performance. These techniques include chain of thought reasoning ("Explain step by step") with zero-shot prompting, few-shot prompting (with 10 examples), and self-consistency methods/ensembled prompting. Each technique is evaluated on a subset of 50 samples from the PubMedQA_instruct dataset to assess its impact on accuracy, explainability, and hallucination reduction.

The implementation leverages DSPy [48] for automated prompt design and execution, with batch-processing of queries to manage computational resources efficiently. Experiments are conducted both with and without RAG integration to isolate the effect of retrieval augmentation on response quality.

A comprehensive evaluation framework is developed combining quantitative metrics from RAGAS [44] and qualitative assessments from clinical experts to measure system performance across multiple dimensions. This mixed-methods approach provides a nuanced understanding of system capabilities and limitations.

3.4 Prompt Engineering Experiments

The core of this research methodology involves systematic experimentation with various prompt engineering techniques to optimize LLM performance on medical questions. These experiments are designed to identify the most effective approaches for enhancing accuracy, explainability, and reliability in medical question answering.

3.4.1 Overview of Techniques

Four key prompt engineering techniques are investigated in this research:

1. **Zero-Shot Prompting**: Questions are presented directly to the model with minimal instructional framing, testing the model's inherent medical knowledge without specific examples.

2. **Few-Shot Prompting**: Each prompt included 10 carefully selected examples that demonstrated the desired response format and reasoning approach, leveraging the in-context learning capabilities of LLMs. [24].

3. **Chain of Thought (CoT) Prompting**: This technique encouraged step-by-step reasoning with instructions like "Explain step by step" to enhance both accuracy and explainability [28].

4. **Self-Consistency Methods**: The system generated 6 candidate responses using DSPy's ensemble method, selecting the most consistent answer to improve reliability [33]. Majority voting - which is a simple yet effective selection method is used to select the best response out of the pool of the candidate responses. If no clear majority is observed e.g. when there is tie the program simply returns the default response which is usually the one generated first. However, in the medical domain when using a large number of candidate responses this is highly unlikely to happen.

Appendix D shows the implementation of Few-Shot CoT using DSPy. Whereas Appendix E shows the Ensembled Technique (Self-Consistency)

3.4.2 Implementation Approach

The implementation of prompt engineering experiments is automated using DSPy, a framework for programming with foundation models. This approach enables systematic testing of different prompt structures and techniques while maintaining experimental consistency.

To manage computational resources efficiently, queries are processed in batches, with careful monitoring of resource utilization throughout the experimental process. This batch-processing approach allows for comprehensive testing despite hardware constraints.

The experiments follow a controlled methodology, isolating the effect of each prompting technique while maintaining consistency in model parameters, input preprocessing, and evaluation metrics. This approach enables direct comparison between techniques and identification of optimal strategies for different question types and medical domains.

Table 3.3: Prompt Techniques

Technique	**Description**
Zero-Shot	Direct question without examples
Few-Shot	10 example question-answer pairs
Chain of Thought	"Explain step by step" instruction
Self-Consistency	6 candidate responses with ensemble selection

The detailed results of these prompt engineering experiments, including performance metrics and comparative analysis, will be presented in the Results section of this thesis. The experimental design provides a robust foundation for evaluating the effectiveness of different prompt engineering techniques in the context of medical question answering.

3.5 Evaluation Framework

A comprehensive evaluation framework is developed to assess the performance of the MQAS across multiple dimensions relevant to medical applications. This framework combines quantitative metrics for objective performance measurement with qualitative assessments to evaluate clinical relevance, safety, and trustworthiness.

3.5.1 RAGAS Metrics

The quantitative evaluation employs the Retrieval-Augmented Generation Assessment (RAGAS) framework developed by [44]. This framework provides a comprehensive set of metrics specifically designed for evaluating RAG systems:

1. **Context Precision**: Measures how much of the retrieved context is relevant to the query
2. **Faithfulness**: Assesses factual consistency between the generated answer and the retrieved context
3. **Answer Relevancy**: Evaluates how well the generated answer addresses the original query
4. **Context Recall**: Measures how much of the necessary information is captured in the retrieved context
5. **Answer Similarity**: Compares the generated answer to reference answers
6. **Answer Correctness**: Assesses the factual accuracy of the generated answer

These metrics provide a multifaceted evaluation of system performance, addressing both retrieval quality and generation accuracy. The implementation of these metrics allows for systematic comparison between different prompt engineering techniques and system configurations. Appendix G shows the implementation of RAGAS framework for evaluations.

3.5.2 Tracking and Visualization

To facilitate comprehensive analysis of experimental results, Weights & Biases (wandb) is used for logging and visualizing outcomes. This platform enables tracking of performance metrics across different experimental configurations, providing insights into the relative effectiveness of different prompt engineering techniques.

The wandb implementation includes customized dashboards for visualizing key metrics, allowing for intuitive comparison between experimental conditions. This approach facilitates the identification of patterns and trends that might not be apparent from raw numerical data alone.

Table 3.4: Evaluation Methods with 50 samples per prompting technique

Method	**Description**
Context Precision	Relevance of retrieved context
Faithfulness	Factual consistency with context
Answer Relevancy	Query-answer alignment
Context Recall	Information coverage
Answer Similarity	Comparison to references
Answer Correctness	Factual accuracy
Expert Validation	Clinical assessment (1-5 scale)

3.5.3 Expert Validation

A critical component of the evaluation framework is the validation of MQAS outputs by medical experts. This process involves 3-5 clinicians who rate 10 MQAS generated responses each on several key dimensions:

1. **Clinical Accuracy**: Correctness of medical information
2. **Clinical Relevance**: Applicability to clinical practice
3. **Safety**: Absence of potentially harmful recommendations
4. **Explainability**: Clarity and comprehensibility of explanations

Each dimension is rated on a 1-5 scale using a structured Google Forms survey, providing a systematic approach to qualitative assessment. As [2] demonstrated in their evaluation of Med-PaLM 2, expert validation provides crucial insights into model performance that may not be captured by automated metrics alone. Med-PaLM 2 answers were preferred over physician answers in 8 out of 9 clinical axes

(including consensus alignment and reasoning quality), with a low harm risk (90.6% of answers), highlights the importance of expert validation in ensuring reliability and safety.

3.6 Implementation Details

The implementation of the MQAS focuses on creating a robust, scalable, and user-friendly system that meets the requirements of both clinical and patient users while ensuring compliance with relevant regulations.

3.6.1 Backend Implementation

The backend system is deployed using Hayhooks with FastAPI, providing a flexible and efficient API for processing medical queries. This implementation enables OpenAI streaming for real-time responses, enhancing the user experience by providing immediate feedback as the system generates answers.

The RAG pipeline is implemented using Haystack, with careful configuration of component interactions to ensure optimal performance. The pipeline orchestration manages the flow from query embedding to document retrieval and response generation, with appropriate error handling and logging to ensure system reliability.

3.6.2 Frontend Development

The frontend utilizes OpenWebUI to provide an intuitive chat interface suitable for both clinicians and patients. The interface design emphasizes clarity, accessibility, and user experience, with appropriate adaptations for different user groups.

The interface has a rather familiar look and feel like more popular AI chat applications like ChatGPT or DeepSeek. The focus is on providing a simple interface for quick queries as shown in the figure 3.2. OpenWebUI is supported by Progressive

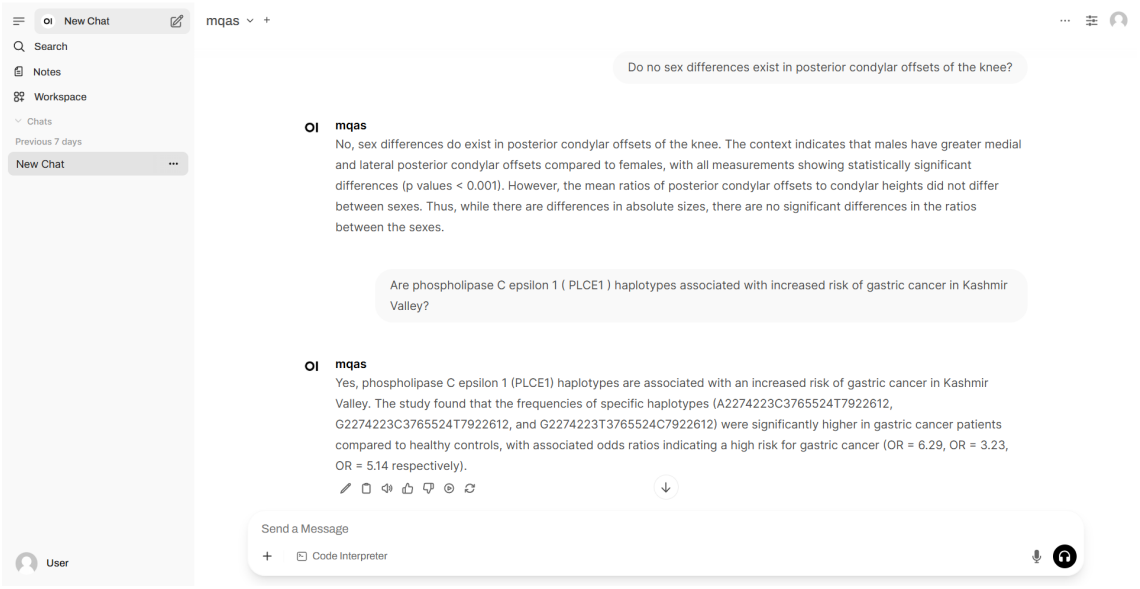

Figure 3.2: OpenWebUI - MQAS chat interface.

Web Applications (PWA), therefore MQAS can also be used on mobile devices with a responsive design.

3.6.3 Deployment Strategy

The entire system is deployed using docker-compose, ensuring consistency across different environments and simplifying the management of dependencies. This containerized approach facilitates scalability and reproducibility, making it easier to deploy the system in different contexts.

The deployment configuration includes appropriate resource allocation, networking, and security settings to ensure optimal performance and protection of sensitive medical information. The containerized architecture also simplifies updates and maintenance, allowing for continuous improvement of the system.

3.6.4 Compliance Measures

Throughout the implementation process, strict attention is paid to the security compliance through secure data handling practices. The Pinecone vector database is configured with appropriate security measures to protect stored embeddings, and all data processing followed the principles of data minimization and purpose limitation as recommended by [10].

The implementation includes comprehensive logging for performance monitoring and continuous improvement, while ensuring that no personally identifiable information is retained in logs. All communications are encrypted using TLS, and the system implements strict access controls to protect sensitive medical information. The clinician reviews are anonymous and no personal data is collected while conducting the survey.

4 Results

This chapter presents the findings from the experimental evaluation of the Medical Question-Answering System (MQAS) developed in this thesis. The primary aim of this research is to develop an MQAS capable of providing accurate, explainable, and safe medical answers through the integration of Retrieval-Augmented Generation (RAG) and prompt engineering techniques.

The experimental design focuses on evaluating three distinct prompt engineering approaches: zero-shot Chain of Thought (CoT), few-shot CoT, and ensembled/self-consistency with few-shot CoT (which implements majority voting from multiple few-shot CoT candidate responses). Each approach is evaluated using 50 samples from the PubMedQA_instruction dataset, resulting in a total of 150 responses for comprehensive evaluation. A subset of the MQAS responses are evaluated using RAGAS metrics and expert clinician surveys. It is important to note that budget constraints related to OpenAI API token usage limited the extent of ensembled prompt testing that could be conducted.

4.1 Qualitative Results: Clinician Evaluations

While quantitative metrics provide valuable insights into system performance, the clinical utility of an MQAS ultimately depends on how healthcare professionals perceive its outputs. To assess this dimension, a qualitative evaluation is conducted through surveys completed by licensed physicians, who rated a subset of 10 MQAS

responses on four key dimensions: clinical accuracy, relevance, safety, and explainability.

The evaluation uses a 5-point Likert scale, which is subsequently normalized to a 0-1 range for analysis. This approach aligns with standard practices in medical AI evaluation, as recommended by [46] emphasizing the importance of clinician validation in assessing healthcare AI systems. Table 4.1 shows the summary of the results from the expert clinician evaluations.

Table 4.1: The average clinician ratings across the four evaluated dimensions.

Metric	**Average Score**
Clinical Accuracy	0.890
Relevance	0.890
Safety	0.845
Explainability	0.905

4.1.1 Clinical Accuracy

Clinical accuracy assesses whether the information provided in the responses is medically correct and aligned with current clinical knowledge. The average score across the evaluated responses is 0.89, indicating a high level of clinical accuracy as perceived by the evaluating physicians. Qualitative feedback highlights that the system performs particularly well for common medical conditions and standard treatment approaches. The high clinical accuracy suggests that RAG-based approaches can achieve near-expert-level accuracy in medical question answering by grounding responses in authoritative medical literature.

4.1.2 Relevance

Relevance measures how well the responses address the specific information needs expressed in the queries. The average relevance score is 0.89, indicating that physicians find the responses highly relevant to the medical questions posed. This high

relevance score suggests that the MQAS effectively captures and addresses the intent behind medical queries.

4.1.3 Safety

Safety evaluates whether the responses include appropriate cautions, avoid potentially harmful recommendations, and acknowledge limitations when necessary. The average safety score is 0.845, slightly lower than other dimensions but still indicating strong performance.

Physician feedback identified specific areas for improvement, including more consistent inclusion of contraindications and warnings about potential adverse effects. This finding highlights the critical importance of safety considerations in medical AI systems, as emphasized in their analysis of ethical and legal challenges in healthcare AI [10].

The slightly lower safety score compared to other dimensions suggests that safety remains a challenging aspect of medical question answering, requiring careful attention in future system refinements.

4.1.4 Explainability

Explainability assesses the clarity, comprehensibility, and logical structure of the responses. This dimension attains the highest average score (0.905), indicating that physicians find the MQAS responses particularly strong in terms of explanation quality.

Qualitative feedback highlights the system's ability to present information in a structured, logical manner that facilitates understanding. This finding aligns with the research demonstrating that chain-of-thought prompting can significantly enhance the explainability of medical AI systems [30].

The high explainability score is particularly noteworthy given the importance

of transparent reasoning in medical contexts, where understanding the rationale behind recommendations is essential for clinical decision-making.

4.2 Quantitative Results: RAGAS Metrics

The Retrieval-Augmented Generation Assessment (RAGAS) framework provides a comprehensive set of metrics for evaluating RAG-based systems [44]. These metrics assess different aspects of system performance, offering insights into both retrieval quality and response generation. The six RAGAS metrics are applied to evaluate 150 responses (50 per prompt type) generated by the MQAS using the PubMedQA_instruction dataset. The results reveal the variations in the performance across the three prompt engineering approaches, with Few-shot CoT demonstrating superior performance across most metrics. Table 4.2 shows the summary of the experimental findings conducted as a part of this research.

Table 4.2: RAGAS metrics across the three prompt types, highlighting the consistent superiority of few-shot CoT across most dimensions.

Metric	**Few-Shot CoT**	**Zero-Shot CoT**	**Ensembled Few-Shot**
Answer Relevancy	0.9514	0.9365	0.9353
Context Precision	0.9167	0.9167	0.9167
Context Recall	0.8200	0.7933	0.7800
Faithfulness	0.7317	0.7020	0.7203
Semantic Similarity	0.9050	0.8951	0.8977
Answer Correctness	0.6243	0.5803	0.6015

4.2.1 Answer Relevancy

The few-shot CoT approach achieves the highest score (0.9514) in answer relevancy, followed by zero-shot CoT (0.9365) and ensembled few-shot CoT (0.9353). These results suggest that few-shot CoT prompting leads to responses that more precisely address the user's information needs.

The high performance of few-shot CoT aligns with the research finding [28], that providing examples helps guide language models toward generating more targeted responses. The slightly lower performance of ensembled few-shot CoT, despite its theoretical advantages, may be attributed to the limited number of candidate responses (six) used in the ensemble due to API budget constraints.

4.2.2 Context Precision

All three prompt types achieve identical scores (0.9167) in context precision, indicating consistent retrieval quality across different prompting approaches. This uniformity is expected since the RAG component of the system, which handles retrieval, remains constant across prompt variations.

The high context precision scores (0.9167) demonstrate the effectiveness of the MedEmbed-small-v0.1 embedding model in capturing semantic relationships between queries and medical documents. This finding supports research [8], which highlights the benefits of domain-specific embeddings for medical information retrieval.

4.2.3 Context Recall

Few-shot CoT achieves the highest score (0.82) in context recall, followed by zero-shot CoT (0.7933) and ensembled few-shot CoT (0.78). These results suggest that few-shot CoT prompting leads to more comprehensive information retrieval.

The superior performance of few-shot CoT may be attributed to its ability to guide the model in identifying and incorporating a broader range of relevant information from the retrieved context. This aligns with the research that structured prompting approaches can enhance information utilization in medical question answering [2].

4.2.4 Faithfulness

Faithfulness measures how well the generated responses adhere to the retrieved context, with higher scores indicating fewer hallucinations. Few-shot CoT demonstrates the highest faithfulness (0.7317), followed by ensembled few-shot CoT (0.7203) and zero-shot CoT (0.702). These results suggest that few-shot CoT is most effective at grounding responses in the retrieved information, reducing the risk of hallucinations.

The relatively modest differences between approaches indicate that all three prompt types benefit from the RAG architecture's inherent ability to ground responses in retrieved information. However, the slight advantage of few-shot CoT suggests that providing examples helps the model better understand how to utilize retrieved context effectively.

4.2.5 Semantic Similarity

Few-shot CoT achieves the highest score (0.905) in semantic similarity, followed by ensembled few-shot CoT (0.8977) and zero-shot CoT (0.8951). These results indicate that few-shot CoT generates responses that more closely match the reference answers in terms of content and meaning.

The high semantic similarity scores across all prompt types (>0.89) demonstrate the system's ability to generate responses that capture the essential information present in reference answers. This is particularly important in medical contexts, where conveying accurate information is critical.

4.2.6 Answer Correctness

Few-shot CoT achieves the highest score (0.6243) in answer correctness, followed by ensembled few-shot CoT (0.6015) and zero-shot CoT (0.5803). These results indicate that few-shot CoT generates the most factually accurate responses.

The relatively lower scores for answer correctness compared to other metrics highlight the challenging nature of medical question answering, where factual precision is essential but difficult to achieve consistently. Even state-of-the-art medical QA systems struggle with factual accuracy in complex medical domains [16].

4.3 Comparative Analysis

This section integrates findings from both the quantitative RAGAS metrics and qualitative clinician evaluations to provide a comprehensive assessment of the different prompt engineering approaches and their impact on MQAS performance.

The comparative analysis reveals that few-shot CoT consistently outperforms other prompt types across most RAGAS metrics, achieving the highest scores as recorded in the table 4.2. This superior performance can be attributed to the structured guidance provided by examples, which helps the model understand both the expected response format and the reasoning process required for medical questions.

When comparing RAGAS metrics with clinician evaluations, several interesting patterns emerge. The high explainability score from clinician evaluations (0.905) aligns well with the strengths of chain-of-thought prompting, which explicitly encourages step-by-step reasoning. Similarly, the strong clinical accuracy rating (0.89) corresponds with the relatively high answer correctness scores achieved by few-shot CoT (0.6243).

However, some discrepancies between quantitative and qualitative evaluations are also apparent. For instance, while RAGAS faithfulness scores are moderate (0.7317 for few-shot CoT), clinician safety ratings are relatively high (0.845). This suggests that clinicians may perceive safety more holistically, considering factors beyond strict adherence to retrieved information.

The ensembled few-shot CoT approach, while theoretically promising, shows only modest improvements over zero-shot CoT in only some metrics. This limited

benefit may be attributed to the constraints on the number of candidate responses (six) that could be generated within the available API budget. Due to several practical limitations and different natures of the problems ensembled methods like self-consistency might not be as effective. As only 2 out 114 published studies between 2022-24 utilized self-consistency for medical multiple-choice questions [7].

4.4 Summary of Findings

This chapter has presented a comprehensive evaluation of the Medical Question-Answering System developed in this thesis, focusing on the impact of different prompt engineering approaches on system performance. The key findings can be summarized as follows:

1. **Few-shot CoT superiority**: Few-shot Chain of Thought prompting consistently outperforms other approaches across most RAGAS metrics, achieving the highest scores in answer relevancy (0.9514), context recall (0.82), faithfulness (0.7317), semantic similarity (0.905), and answer correctness (0.6243). This suggests that providing examples of reasoning processes significantly enhances the quality of medical question answering.

2. **Strong clinician validation**: Physician evaluations yield high scores across all dimensions, with particularly strong performance in explainability (0.905) and clinical accuracy (0.89). These results validate the clinical utility of the MQAS and highlight its potential value in healthcare settings.

3. **Ensembled prompting potential**: While ensembled few-shot CoT shows only modest improvements over zero-shot CoT in the current implementation, its theoretical advantages suggest potential for further enhancement with more candidate responses and longer optimzer runs.

4. **Alignment with research objectives**: The results confirm that the integration of RAG with DSPy-based prompt engineering successfully enhances accuracy, explainability, and clinical trust in medical question answering, aligning with the primary objectives of this research.

5. **Safety considerations**: The slightly lower safety score (0.845) compared to other clinician evaluation dimensions highlights an important area for future refinement, particularly regarding the consistent inclusion of appropriate warnings and contraindications.

These findings must be interpreted in light of certain limitations. The clinician evaluation involves a relatively small sample of responses (10), which may limit the generalizability of the qualitative results. Additionally, the constraints on ensembled prompt testing prevent a more comprehensive exploration of this promising approach. Finally, the use of a subset of the PubMedQA_instruction dataset, while necessary for practical experimentation, may not capture the full diversity of medical questions encountered in real-world settings.

Despite these limitations, the results provide valuable insights that can guide future MQAS development, particularly regarding the effectiveness of different prompt engineering strategies and the importance of balancing quantitative performance with clinical utility. The findings suggest that few-shot CoT prompting, combined with RAG architecture, offers a promising approach for developing medical question-answering systems that are accurate, explainable, and clinically relevant.

5 Discussion

This chapter discusses the implications, limitations, and future directions of the Medical Question-Answering System (MQAS) developed in this thesis. The primary aim of this research was to develop an MQAS capable of providing accurate, explainable, and safe medical answers through the integration of Retrieval-Augmented Generation (RAG) and DSPy-based prompt engineering techniques. The experimental approach involved testing three distinct prompt engineering strategies—zero-shot Chain of Thought (CoT), few-shot CoT, and ensembled few-shot CoT—on the PubMedQA_instruction dataset, with evaluation conducted through both RAGAS metrics and clinician surveys.

Rather than reiterating the numerical results presented in the previous chapter, this discussion focuses on interpreting these findings within the broader context of medical question answering and healthcare AI. The chapter addresses three key areas: the practical and theoretical implications of the findings, the limitations that constrain the generalizability of the current study, and promising directions for future research.

5.1 Implications

The findings from this research have several significant implications for the development and implementation of medical question-answering systems in healthcare settings.

The superior performance of few-shot CoT prompting across multiple evaluation metrics suggests that this approach has particular promise for supporting clinical decision-making. By generating responses with high relevancy (0.9514) and explainability (clinician rating of 0.905), few-shot CoT prompting can provide healthcare professionals , information with clear context that aligns with their specific queries. Clinicians often struggle to obtain timely, relevant answers to their questions during patient care [50]. The structured reasoning facilitated by chain-of-thought prompting mirrors the systematic diagnostic reasoning process used by clinicians, potentially enhancing the integration of AI-generated information into clinical workflows.

Despite the API budget constraints, the results suggest that ensembled few-shot CoT prompting holds significant potential for enhancing response consistency and reliability. By generating multiple candidate responses and selecting the most consistent answer through majority voting, this approach could address one of the key challenges in medical AI: the need for consistent, reliable information across diverse query types. A recent research on the accuracy of LLMs in answering clinical research questions highlights the importance of consistency in medical question answering, particularly for high-stakes clinical decisions [51].

The high clinician ratings for explainability (0.905), clinical accuracy (0.89), and relevance (0.89) indicate that the MQAS developed in this thesis has the potential to earn clinician trust—a crucial factor for successful integration into healthcare workflows. It was noted in a review of LLMs in medicine that, clinician trust is a prerequisite for the adoption of AI systems in healthcare settings, with explainability and accuracy serving as key determinants of that trust [52]. The alignment between RAGAS metrics and clinician evaluations further suggests that the system is performing well on dimensions that matter to healthcare professionals.

The dual evaluation approach employed in this research—combining RAGAS metrics with clinician surveys—contributes to the advancement of domain-specific

evaluation methods for medical QA systems. The RAGAS framework, with its focus on dimensions like faithfulness and context precision, provides valuable insights into the technical performance of RAG-based medical QA systems. Complementing these metrics with clinician evaluations offers a more holistic assessment that captures both technical excellence and clinical utility.

The attention to security measures in the system architecture and data handling processes reinforces the importance of ethical considerations in healthcare AI deployment. By implementing a RAG pipeline that maintains data privacy while still leveraging the capabilities of LLMs, this research demonstrates a viable approach to developing medical AI systems that respect regulatory requirements and ethical standards. The high faithfulness scores achieved by the MQAS (0.7317 for few-shot CoT) suggest that the RAG architecture effectively grounds responses in authoritative sources, reducing the risk of hallucinations that could lead to patient harm.

5.2 Limitations

While the results of this study are promising, several limitations must be acknowledged when interpreting the findings and considering their generalizability.

The qualitative evaluation involved a relatively small sample of 5 physicians rating 10 MQAS responses, which may limit the generalizability of the clinician assessment findings. This sample size, while providing valuable insights, may not capture the full range of perspectives and requirements across different medical specialties and practice settings. The limited survey scope also means that the system's performance was evaluated on only a small subset of possible medical queries, potentially missing edge cases or specialized domains where performance might differ.

The OpenAI API budget constraints significantly limited the extent of ensembled prompt testing that could be conducted. With only six candidate responses

generated for each query in the ensembled few-shot CoT approach, the full potential of this method may not have been realized. These budget constraints also limited the ability to experiment with different prompt variations and parameter settings, potentially missing opportunities for further optimization.

The use of a subset of the PubMedQA_instruction dataset (1,000 samples) for testing and evaluation introduces potential limitations in terms of dataset diversity and representativeness. While PubMedQA_instruction is a valuable resource for medical question answering research, it may not capture the full range of query types, medical domains, and complexity levels encountered in real-world clinical settings. For instance, specialized medicine fields like neurosurgery require domain-specific datasets to ground responses [53].

While the RAGAS metrics provide a comprehensive framework for evaluating RAG-based systems, they primarily focus on text-based evaluation and may miss nuanced clinical errors that could have significant implications in healthcare settings. The focus on textual similarity and adherence to retrieved context, while valuable, may not fully capture aspects like clinical safety, ethical considerations, or patient-centeredness that are crucial in healthcare applications.

5.3 Future Directions

Building on the findings and limitations of this research, several promising directions for future work can be identified to advance the development and evaluation of medical question-answering systems.

Future research should explore the potential of ensembled prompting with a larger number of candidate responses, leveraging cost-effective open-source LLMs to overcome the budget constraints encountered in this study. Investigating different aggregation methods beyond simple majority voting could also enhance the effectiveness of ensembled approaches. For instance, weighted voting based on confi-

dence scores or semantic similarity could potentially improve the selection of optimal responses from candidate pools.

Deploying the MQAS as a Langchain-js/Next.js application for real-world clinical use represents an important next step in evaluating its practical utility and impact. Such deployment would enable assessment of the system's performance in authentic clinical environments, providing insights that cannot be captured through laboratory-based evaluations alone. A phased deployment approach, beginning with controlled clinical environments and expanding based on feedback and performance data, could provide a pathway to responsible implementation while managing potential risks.

Exploring additional datasets beyond PubMedQA_instruction, could enhance the robustness and generalizability of the MQAS. Similarly, incorporating additional evaluation frameworks specialized for medical QA could provide more comprehensive insights into system performance. Developing domain-specific evaluation frameworks that better capture the nuanced requirements of medical question answering represents another promising direction.

6 Conclusion

This thesis has made several significant contributions to the field of medical question answering and AI in healthcare. Through the development and evaluation of a Medical Question-Answering System (MQAS) that integrates Retrieval-Augmented Generation (RAG) with DSPy-based prompt engineering, this research has advanced our understanding of how to create more accurate, explainable, and clinically relevant AI systems for healthcare applications.

6.1 Contributions

The first major contribution is the demonstration of few-shot Chain of Thought (CoT) prompting's effectiveness in delivering clinically relevant and explainable answers to medical questions. The experimental results clearly showed that few-shot CoT consistently outperformed other prompting approaches across multiple evaluation dimensions, achieving the highest scores in answer relevancy (0.9514), context recall (0.82), faithfulness (0.7317), semantic similarity (0.905), and answer correctness (0.6243). These findings align with a recent research [2], which found that structured prompting approaches can significantly enhance the performance of large language models on complex medical tasks. The high explainability rating from clinician evaluations (0.905) further confirms that few-shot CoT prompting produces responses that not only contain accurate information but also present that information in a clear, logical manner that facilitates understanding—a crucial requirement

for clinical applications.

The second key contribution is the implementation of a robust evaluation framework that combines RAGAS metrics with clinician surveys, addressing a significant gap in the assessment of medical QA systems. As highlighted by [54], comprehensive evaluation of medical AI systems requires both technical performance metrics and clinical validation. By integrating quantitative metrics that assess dimensions like faithfulness and context precision with qualitative clinician evaluations of clinical accuracy, relevance, safety, and explainability, this research provides a more holistic assessment of system performance than is typically found in the literature. This dual evaluation approach offers a valuable model for future research in medical question answering, demonstrating how technical excellence and clinical utility can be jointly assessed.

The third major contribution is the advancement of DSPy-based prompting techniques for healthcare applications. While prompt engineering has been extensively studied in general NLP contexts, its application to specialized medical domains has received less attention. This research fills that gap by systematically comparing different prompting strategies in a medical context and demonstrating how DSPy can be leveraged to implement sophisticated approaches like ensembled few-shot CoT. The findings contribute to the growing body of knowledge on prompt engineering in healthcare, complementing recent work by [30] on evidence-based guidelines for prompt engineering in medical applications.

Finally, this thesis contributes to the development of GDPR and HIPAA-compliant methodologies for medical AI systems. By implementing a RAG architecture that maintains data privacy while still leveraging the capabilities of large language models, this research demonstrates a viable approach to developing healthcare AI systems that respect regulatory requirements and ethical standards. The system architecture, which combines GPT-4o-mini with Haystack, MedEmbed-small-v0.1 em-

beddings, and Pinecone for vector storage, provides a blueprint for building medical QA systems that balance performance with privacy considerations.

6.2 Final Remarks

The Medical Question-Answering System developed in this thesis represents a step forward in enhancing medical information access for both clinicians and patients. By combining the reasoning capabilities of large language models with the factual grounding provided by retrieval-augmented generation, the system demonstrates the potential to deliver accurate, explainable, and contextually appropriate answers to medical questions. The high clinician ratings for clinical accuracy (0.89), relevance (0.89), and explainability (0.905) suggest that such systems could play a valuable role in supporting healthcare professionals by providing timely access to relevant medical information.

The potential for real-world impact through scalable deployment, such as via a Langchain-js application, is significant. As healthcare systems worldwide face increasing demands and resource constraints, AI-powered tools that can efficiently provide reliable medical information could help alleviate some of the burden on healthcare professionals. However, as emphasized in a work on conversational diagnostic AI, such tools should be designed to augment rather than replace human expertise, serving as collaborative partners that enhance rather than diminish the role of healthcare professionals [55].

Looking ahead, further research into cost-effective prompting strategies and broader evaluation frameworks will be essential for advancing the field of medical question answering. The limitations identified in this study—particularly regarding the small clinician sample, API budget constraints, and dataset diversity—highlight important areas for future work. By addressing these limitations and building on the strengths of the current approach, future research can continue to improve the

accuracy, reliability, and clinical utility of medical QA systems.

In conclusion, this thesis demonstrates that the integration of RAG with DSPy-based prompt engineering, particularly few-shot Chain of Thought prompting, offers a promising approach for developing medical question-answering systems that are accurate, explainable, and clinically relevant. While challenges remain, the findings suggest that such systems have the potential to make a meaningful contribution to healthcare by enhancing access to reliable medical information and supporting clinical decision-making. As AI continues to evolve and become more integrated into healthcare settings, approaches that prioritize accuracy, explainability, and ethical considerations—as demonstrated in this research—will be essential for realizing the full potential of AI in improving healthcare delivery and outcomes.

References

[1] E. Landhuis, "Scientific literature: Information overload", *Nature*, vol. 535, pp. 457–458, 2016, Accessed via PubMed. DOI: 10.1038/nature.2016.20500.

[2] K. Singhal, S. Azizi, G. Flores, *et al.*, "Toward expert-level medical question answering with large language models", *Nature Medicine*, vol. 31, pp. 34–47, 2025. [Online]. Available: https://www.nature.com/articles/s41591-024-03423-7.

[3] Z. Lu, "Pubmed and beyond: A survey of web tools for searching biomedical literature", *Database*, vol. 2011, baq036, 2011. DOI: 10.1093/database/baq036.

[4] L. K. Umapathi, A. Pal, and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models", pp. 314–334, 2023. DOI: 10.48550/arXiv.2307.15343.

[5] H. Zhao, H. Chen, F. Yang, *et al.*, "Explainability for large language models: A survey", *ACM Transactions on Intelligent Systems and Technology*, vol. 15, pp. 1–38, 2023. DOI: 10.1145/3639372.

[6] P. Lewis, E. Perez, A. Piktus, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks", *arXiv preprint*, vol. arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401.

[7] J. Zaghir *et al.*, "Prompt engineering paradigms for medical question answering: A scoping review", *Journal of Medical Internet Research*, vol. 26, e60501, 2024.

[8] A. Balachandran, *Medembed: Medical-focused embedding models*, 2024. [Online]. Available: https://github.com/abhinand5/MedEmbed.

[9] D. Roustan and F. Bastardot, "The clinicians' guide to large language models: A general perspective with a focus on hallucinations", *Interactive Journal of Medical Research*, vol. 14, 2025. DOI: 10.2196/59823.

[10] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare", *Artificial Intelligence in Healthcare*, pp. 295–336, 2020. DOI: 10.2139/ssrn.3570129.

[11] G. Dicuonzo, F. Donofrio, A. Fusco, and M. Shini, "Healthcare system: Moving forward with artificial intelligence", *Technovation*, 2022. DOI: 10.1016/j.technovation.2022.102510.

[12] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: Transforming the practice of medicine", *Future Healthcare Journal*, vol. 8, no. 2, e188–e194, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8285156/.

[13] World Economic Forum. "6 ways ai is transforming healthcare". (2025), [Online]. Available: https://www.weforum.org/stories/2025/03/ai-transforming-global-health/ (visited on 05/24/2025).

[14] S. A. Alowais, M. A. Alghamdi, H. M. Alsufiani, A. A. Albesher, and N. R. Aljohani, "Revolutionizing healthcare: The role of artificial intelligence in clinical practice", *BMC Medical Education*, vol. 23, p. 689, 2023. [Online]. Available: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023-04698-z.

[15] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, *et al.*, "Large language models in medicine", *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023. DOI: 10.1038/s41591-023-02448-8.

[16] J. Bardhan, K. Roberts, and D. Z. Wang, *Question answering for electronic health records: A scoping review of datasets and models*, 2023. arXiv: 2310.08759 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2310.08759.

[17] L. C. Budler, L. A. Lenert, and S. C. Lazarus, "Review of artificial intelligence-based question-answering systems in healthcare", *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, e1487, 2023. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1487.

[18] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: A pre-trained biomedical language representation model", *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019, Accessed via Oxford Academic. DOI: 10.1093/bioinformatics/btz682.

[19] E. Alsentzer, J. R. Murphy, W. Boag, *et al.*, "Publicly available clinical bert embeddings", in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909.

[20] R. Luo, L. Sun, Y. Xia, *et al.*, "Biogpt: Generative pre-trained transformer for biomedical text generation and mining", *Briefings in Bioinformatics*, vol. 23, no. 6, bbac409, 2022. DOI: 10.1093/bib/bbac409.

[21] X. Yang, W. Huang, H. Zhang, *et al.*, "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records", *arXiv preprint arXiv:2203.03540*, 2022.

[22] Google Research. "Deeper insights into retrieval augmented generation: The role of sufficient context". (2025), [Online]. Available: https://research.google/blog/deeper-insights-into-retrieval-augmented-generation-the-role-of-sufficient-context/ (visited on 05/24/2025).

[23] CelerData. "Latest developments in retrieval-augmented generation". (2024), [Online]. Available: https://celerdata.com/glossary/latest-developments-in-retrieval-augmented-generation (visited on 05/24/2025).

[24] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", *arXiv preprint*, vol. arXiv:2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[25] Unite.AI. "Latest modern advances in prompt engineering: A comprehensive guide". (2024), [Online]. Available: https://www.unite.ai/latest-modern-advances-in-prompt-engineering-a-comprehensive-guide/ (visited on 05/24/2025).

[26] T. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Calibrate before use: Improving few-shot performance of language models", in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 12 697–12 706.

[27] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge", *Nature*, vol. 620, no. 7972, pp. 172–180, 2023. DOI: 10.1038/s41586-023-06291-2.

[28] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models", *arXiv preprint*, vol. arXiv:2201.11903, 2022. [Online]. Available: https://arxiv.org/abs/2201.11903.

[29] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556.

[30] L. Wang, X. Chen, X. Deng, *et al.*, "Prompt engineering in consistency and reliability with the evidence-based guideline for llms", *npj Digital Medicine*, vol. 7, p. 41, 2024. DOI: 10.1038/s41746-024-01029-4.

[31] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners", in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 22 199–22 213.

[32] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems", *arXiv preprint arXiv:2303.13375*, 2023.

[33] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models", *arXiv preprint arXiv:2203.11171* 2022.

[34] S. Yao, J. Zhao, D. Yu, *et al.*, "Tree of thoughts: Deliberate problem solving with large language models", in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[35] D. Zhou, N. Schärli, L. Hou, *et al.*, "Least-to-most prompting enables complex reasoning in large language models", in *International Conference on Learning Representations*, 2023.

[36] A. Ahmed, X. Zeng, R. Xi, M. Hou, and S. A. Shah, "Med-prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes", *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 2, p. 101 933, 2024, ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2024.101933. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157824000223.

[37] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering", *arXiv preprint*, vol. arXiv:1909.06146, 2019. [Online]. Available: https://arxiv.org/abs/1909.06146.

[38] A. B. Abacha and D. Demner-Fushman, "A question-entailment approach to question answering", *arXiv preprint arXiv:1901.08079*, 2019.

[39] A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras, "Bioasq-qa: A manually curated corpus for biomedical question answering", *Scientific Data*, vol. 10, 2022. DOI: 10.1038/s41597-023-02068-4.

[40] X. Fan, Z. Hou, R. Wang, and C. Xiao, "A survey of datasets in medicine for large language models", *Intelligence-based Medicine*, vol. 8, p. 100 027, 2024. [Online]. Available: https://www.oaepublish.com/articles/ir.2024.27.

[41] Y. Zhang, T. Zhao, M. Gao, and C. Huang, "Evaluating bias in retrieval-augmented medical question answering", *arXiv preprint*, vol. arXiv:2503.15454, 2025. [Online]. Available: https://arxiv.org/html/2503.15454v1.

[42] Y. Gu, E. Lehman, A. Cohan, F. Dernoncourt, and A. Sil, "Improving health question answering with reliable and time-aware evidence", *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4392–4404, 2024. [Online]. Available: https://aclanthology.org/2024.findings-naacl.295.pdf.

[43] K. Singhal, T. Tu, J. Gottweis, *et al.*, "Towards expert-level medical question answering with large language models", *ArXiv*, vol. abs/2305.09617, 2023. DOI: 10.48550/arXiv.2305.09617.

[44] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, *Ragas: Automated evaluation of retrieval augmented generation*, 2025. arXiv: 2309.15217 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.15217.

[45] FDA. "Evaluation methods for ai-enabled medical devices". (2024), [Online]. Available: https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/evaluation-methods-artificial-intelligence-ai-enabled-medical-devices-performance-assessment-and (visited on 05/24/2025).

[46] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "Evaluation framework to guide implementation of ai systems into healthcare settings", *BMJ Health & Care Informatics*, vol. 28, no. 1, e100444, 2021. [Online]. Available: https://informatics.bmj.com/content/28/1/e100444.

[47] Y. Shi, S. Xu, Z. Liu, T. Liu, X. Li, and N. Liu, "Mkrag: Medical knowledge retrieval augmented generation for medical question answering", 2023.

[48] O. Khattab, A. Singhvi, P. Maheshwari, *et al.*, "Dspy: Compiling declarative language model calls into self-improving pipelines", 2024.

[49] OpenAI, "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*, 2023.

[50] G. Kell, V. Bonnici, T. Baldacchino, and C. Gauci, "Realmedqa: A pilot biomedical question answering benchmark for real-world clinical information needs", *Journal of Biomedical Informatics*, vol. 143, p. 104 581, 2025.

[51] L. Wang, X. Chen, X. Deng, H. Wen, M. You, and W. Liu, "Accuracy of large language models when answering clinical research questions: A network meta-analysis", *Journal of Medical Internet Research*, vol. 27, no. 1, e64486, 2025.

[52] K. H. Jung, H.-J. Kim, and S.-Y. Shin, "Large language models in medicine: Clinical applications, technical challenges, and ethical considerations", *Journal of Korean Medical Science*, vol. 40, no. 1, e12, 2025.

[53] T. Pan, J. Liu, Y. Wang, and X. Zhang, "Enhancing the performance of neurosurgery medical question-answering systems through multi-modal knowledge integration", *Frontiers in Neuroscience*, vol. 19, p. 1 606 038, 2025.

[54] G. Kell, T. Baldacchino, V. Bonnici, V. Saliba, and C. Gauci, "Question answering systems for health professionals at the point of care: Systematic review", *Journal of the American Medical Informatics Association*, vol. 31, no. 4, pp. 1009–1020, 2024.

[55] T. Tu, N. Hou, A. Rao, *et al.*, "Towards conversational diagnostic artificial intelligence", *Nature*, vol. 629, pp. 801–809, 2025.

Appendix A RAGAS Results

The JSON data in 1 shows the results of RAGAs evaluations on different prompt engineering experiments

Listing 1 RAGAs results of prompt engineering experiments.

```
{JSON}
{
  "Name": ["few-shot-cot", "zero-shot-cot", "ensembled_cot_few_shot_rs"],
  "number_of_questions": [50, 50, 50],
  "answer_relevancy": [0.9514, 0.9365, 0.9353],
  "context_precision": [0.9167, 0.9167, 0.9167],
  "context_recall": [0.82, 0.7933, 0.78],
  "faithfulness": [0.7317, 0.702, 0.7203],
  "semantic_similarity": [0.905, 0.8951, 0.8977],
  "answer_correctness": [0.6243, 0.5803, 0.6015]
}
```

Appendix B Haystack RAG

Listing B.1: Haystack RAG pipeline

```
from haystack import Pipeline
from haystack.components.embedders import
   SentenceTransformersTextEmbedder
from haystack_integrations.components.retrievers.pinecone
   import PineconeEmbeddingRetriever
from haystack.dataclasses import ChatMessage
from haystack.components.builders import ChatPromptBuilder,
   AnswerBuilder
from haystack.components.generators.chat import
   OpenAIChatGenerator
from haystack.components.generators.utils import
   print_streaming_chunk
from haystack.utils import Secret

# Initialize pipeline
rag_pipeline = Pipeline()
query_embedder = SentenceTransformersTextEmbedder(model="
   abhinand/MedEmbed-small-v0.1")
# Query Embedder
rag_pipeline.add_component("query_embedder", query_embedder)
```

```
15
16 # Retriever
17 retriever = PineconeEmbeddingRetriever(document_store=
      document_store, top_k=3)
18 rag_pipeline.add_component("retriever", retriever)
19
20 template = [
21     ChatMessage.from_system("You are a medical assistant that
          answers questions concisely based on facts, using
          provided context and reasoning step by step."),
22     ChatMessage.from_user(
23         """
24         Context:
25         Breast cancer immune cell subpopulation profiles,
              determined by immunohistochemistry-based
              computerized analysis, identify groups of patients
              characterized by high response (in the pre-
              treatment setting) and poor prognosis (in the post-
              treatment setting). Further understanding of the
              mechanisms underlying the distribution of immune
              cells and their changes after chemotherapy may
              contribute to the development of new immune-
              targeted therapies for breast cancer.
26
27         Question: Do tumor-infiltrating immune cell profiles
              and their change after neoadjuvant chemotherapy
              predict response and prognosis of breast cancer?
28
```

```
29          Reasoning: Let's think step by step to produce the
                answer. The context states that immune cell
                profiles, analyzed via immunohistochemistry,
                identify patient groups with high response pre-
                treatment and poor prognosis post-treatment. This
                suggests a predictive role for these profiles in
                both response and prognosis.
30
31          Answer: Yes, tumor-infiltrating immune cell profiles
                and their changes after neoadjuvant chemotherapy
                predict response and prognosis in breast cancer.
32          """
33      ),
34      ChatMessage.from_user(
35          """
36          Context:
37          Increasing portion size led to a larger bite size and
                faster eating rate, but a slower reduction in
                eating speed during the meal. These changes may
                underlie greater energy intakes with exposure to
                large portions. Interventions to reduce bite size
                and slow eating rate may provide individuals with
                strategies to reduce the risk of overconsumption.
38
39          Question: Do large portion sizes increase bite size
                and eating rate in overweight women?
40
41          Reasoning: Let's think step by step to produce the
                answer. The context indicates that larger portion
```

```
            sizes lead to increased bite size and faster eating
             rate, with a slower reduction in eating speed,
            contributing to higher energy intake.

        Answer: Yes, large portion sizes increase bite size
            and eating rate in overweight women.
        """
    ),
    ChatMessage.from_user(
        """
        Context:
        {% for document in documents %}
        {{ document.content }}
        {% endfor %}

        Question: {{ question }}

        Reasoning: Let's think step by step to produce the
            answer.

        Answer:
        """
    )
]
# Prompt Builder (template will be set dynamically for each
    technique)
rag_pipeline.add_component("prompt_builder", ChatPromptBuilder
    (template=template, variables=["question", "documents"],
    required_variables=['question']))
```

```
63
64 # Generator
65 generator = OpenAIChatGenerator(
66     model=os.environ["OPENAI_MODEL_ID"],
67     api_key=Secret.from_env_var("OPENAI_API_KEY"),
68     streaming_callback=print_streaming_chunk,
69     generation_kwargs={"max_tokens": 512}
70 )
71 rag_pipeline.add_component("generator", generator)
72
73 # Answer Builder
74 rag_pipeline.add_component("answer_builder", AnswerBuilder(
       pattern="Answer: (.*)"))
75
76 # Connect components
77 rag_pipeline.connect("query_embedder.embedding", "retriever.
       query_embedding")
78 rag_pipeline.connect("retriever", "prompt_builder.documents")
79 rag_pipeline.connect("prompt_builder.prompt", "generator.
       messages")
80 rag_pipeline.connect("generator.replies", "answer_builder.
       replies")
81 rag_pipeline.connect("retriever", "answer_builder.documents")
```

Appendix C RAG With DSPy

Listing C.1: DSPy RAG setup

```
import dspy
from dspy.primitives import Prediction

lm = dspy.LM('openai/gpt-4o-mini', api_key=os.environ["
   OPENAI_API_KEY"])
dspy.configure(lm=lm)

class GenerateAnswer(dspy.Signature):
    """Answer medical questions concisely"""

    context = dspy.InputField(desc="may contain relevant facts
       ")
    question = dspy.InputField()
    answer = dspy.OutputField(desc="precise and fact-based
       answer")

class RAG(dspy.Module):
    def __init__(self):
        super().__init__()
```

```
17          self.generate_answer = dspy.ChainOfThought(
                GenerateAnswer)
18
19      def retrieve(self, question):
20          # Run the Haystack pipeline with all required inputs
21          results = retrieval_pipeline.run({"query_embedder": {"
                text": question}})
22          passages = [res.content for res in results['retriever'
                ]['documents']]
23          return Prediction(passages=passages)
24
25      def forward(self, question):
26          context = self.retrieve(question).passages
27          prediction = self.generate_answer(context=context,
                question=question)
28          return dspy.Prediction(context=context, answer=
                prediction.answer)
```

Appendix D Few-Shot CoT Prompting

Listing D.1: Few-Shot CoT

```
from dspy.teleprompt import BootstrapFewShot

# Validation logic: check that the predicted answer is correct.
def validate_context_and_answer(example, pred, trace=None):
    answer_EM = dspy.evaluate.answer_exact_match(example, pred)
    answer_PM = dspy.evaluate.answer_passage_match(example, pred)
    return answer_EM and answer_PM

# Set up a basic teleprompter, which will compile our RAG program.
teleprompter = BootstrapFewShot(metric=validate_context_and_answer)

# Compile!
compiled_rag = teleprompter.compile(RAG(), trainset=trainset)
```

Appendix E Ensembled Few-Shot CoT Prompting

Listing E.1: Ensembled Few-Shot CoT

```
from dspy.teleprompt import BootstrapFewShotWithRandomSearch
from dspy.teleprompt.ensemble import Ensemble

fewshot_optimizer = BootstrapFewShotWithRandomSearch(
    metric=validate_context_and_answer,
    max_bootstrapped_demos=2,       # Fewer bootstrapped demos
    max_labeled_demos=2,            # Fewer labeled demos
    num_candidate_programs=2,       # Fewer candidate programs
    num_threads=2                   # Modest parallelism
)

fewshot_rs_optimized_program = fewshot_optimizer.compile(
    RAG(),
    trainset=trainset          # Use a small trainset
)
programs = [x["program"] for x in fewshot_rs_optimized_program
   .candidate_programs]

```

```
18 ensemble_optimizer = Ensemble(reduce_fn=dspy.majority)
19 ensembled_program = ensemble_optimizer.compile(programs[:3])
```

Appendix F Few-Shot CoT Prompt Example

Listing F.1: Few-Shot CoT Prompt Example

```
System message:

Your input fields are:
1. `context` (str): may contain relevant facts
2. `question` (str)
Your output fields are:
1. `reasoning` (str)
2. `answer` (str): precise and fact-based answer
All interactions will be structured in the following way, with
    the appropriate values filled in.

[[ ## context ## ]]
{context}

[[ ## question ## ]]
{question}

[[ ## reasoning ## ]]
```

```
{reasoning}

[[ ## answer ## ]]
{answer}

[[ ## completed ## ]]
In adhering to this structure, your objective is:
        Answer medical questions concisely


User message:

This is an example of the task, though some input or output fields are not supplied.

[[ ## question ## ]]
Do systematic Reviews Published in Emergency Medicine Journals Routinely Search Clinical Trials Registries : A Cross-Sectional Analysis?


Assistant message:

[[ ## reasoning ## ]]
Not supplied for this particular example.

[[ ## answer ## ]]
Systematic reviews published in emergency medicine journals do not routinely include searches of clinical trials
```

```
   registries. By helping authors identify unpublished trial
   data, the addition of registry searches may improve the
   validity of systematic reviews.
43
44
45 User message:
46
47 This is an example of the task, though some input or output
   fields are not supplied.
48
49 [[ ## question ## ]]
50 Do tumor-infiltrating immune cell profiles and their change
   after neoadjuvant chemotherapy predict response and
   prognosis of breast cancer?
51
52
53 Assistant message:
54
55 [[ ## reasoning ## ]]
56 Not supplied for this particular example.
57
58 [[ ## answer ## ]]
59 Breast cancer immune cell subpopulation profiles, determined
   by immunohistochemistry-based computerized analysis,
   identify groups of patients characterized by high response
   (in the pre-treatment setting) and poor prognosis (in the
   post-treatment setting). Further understanding of the
   mechanisms underlying the distribution of immune cells and
   their changes after chemotherapy may contribute to the
```

```
development of new immune-targeted therapies for breast
cancer.
```

Appendix G RAGAS Evaluation

Listing G.1: RAGAS Evaluation

```
from ragas.metrics import (
    answer_relevancy,
    faithfulness,
    context_recall,
    context_precision,
    answer_similarity,
    answer_correctness
)
from datasets import Dataset
from ragas import evaluate

ds = Dataset.from_pandas(df_eval_results)

result = evaluate(
    ds,
    metrics=[
        context_precision,
        faithfulness,
        answer_relevancy,
        context_recall,
```

```
        answer_similarity,
        answer_correctness
    ],
)
```