# Medical Chatbot: A RAG approach

Diljyot Singh
School of Math and Statistics
University of Guelph
Guelph, Canada
Email: diljyot@uoguelph.ca

Puja Saha
School of Engineeirng
University of Guelph
Guelph, Canada
Email: psaha03@uoguelph.ca

***Abstract*—The use of chatbots in the medical sector is on the rise, aiming to enhance patient care, deliver information, and streamline operations. These chatbots facilitate communication between patients and specialists through various means such as phone calls, video calls, messages, and emails, thereby boosting engagement and allowing medical staff to allocate their time more efficiently. Notable healthcare chatbots like Med-Palm, OneRemission, Ada Health, Florence, and Babylon Health offer assistance in managing patient inquiries and providing personalized care. It's important to note that while chatbots can handle routine queries and offer tailored support, they are intended to complement rather than replace medical professionals. They offer advantages such as round-the-clock availability, reduced wait times, and cost-effectiveness, while also delivering prompt and personalized assistance for basic patient needs.**

**Keywords:** LLM, RAG, Chatbot, Medical, PubMed

## I. Introduction

Large Language Models (LLMs) have significantly advanced natural language processing capabilities, demonstrating remarkable performance across various domains. However, their integration into the medical field presents unique challenges and concerns. In this section, we delve into the limitations of LLMs in healthcare contexts and explore the potential of Retrieval-Augmented Generation (RAG) models as a solution. We then introduce a comprehensive methodology for developing a medical chatbot using RAG, emphasizing its advantages and applications in addressing these challenges.

The adoption of Large Language Models (LLMs) in healthcare settings holds tremendous promise for enhancing clinical decision-making and patient care. However, despite their impressive language processing capabilities, LLMs face several limitations when applied to the medical domain. One significant challenge is the tendency of LLMs to exhibit unwarranted confidence in their responses, even when providing inaccurate or potentially harmful medical advice. This phenomenon, known as overconfidence, poses a significant risk to patient safety and healthcare outcomes, highlighting the need for more robust and contextually aware AI systems in healthcare.

Furthermore, LLMs are susceptible to bias present in the training data, which can lead to the propagation of misinformation and biased recommendations. In the context of healthcare, where decisions can have profound implications for patient well-being, mitigating bias is of utmost importance to ensure equitable and evidence-based care delivery. Additionally, the complexity of medical language and the diverse contexts in which medical information is presented pose challenges for LLMs in accurately comprehending and responding to user queries.

In response to these challenges, Retrieval-Augmented Generation (RAG) models offer a promising approach to developing more robust and contextually aware medical chatbots. By combining retrieval-based techniques with generative models, RAG enables the incorporation of external knowledge sources and contextually relevant responses, thereby enhancing the accuracy and reliability of AI-driven medical assistance. In this paper, we propose a comprehensive methodology for leveraging RAG in the development of medical chatbots, aiming to address the limitations of LLMs and facilitate more effective and personalized healthcare interactions.

## II. Problem Definition

The project aims to develop a robust Retrieval-Augmented Generation (RAG) based medical chatbot to overcome the limitations observed in existing chatbot systems within the medical domain. The inadequacies of current chatbots often manifest in their inability to comprehend the contextual intricacies of medical queries, resulting in the provision of generic or irrelevant responses. Such deficiencies not only hinder effective communication between patients and healthcare providers but also pose risks by potentially disseminating inaccurate or outdated medical advice.

Specific challenges targeted by this project include:

1. Lack of Contextual Understanding: Existing medical chatbots often struggle to grasp the nuanced context of medical queries, leading to responses that may not adequately address the user's concerns.

2. Dissemination of Inaccurate Information: Chatbots lacking robust retrieval and generation mechanisms may inadvertently provide incorrect or outdated medical advice, thereby compromising patient safety.

3. Limited Access to Relevant Data: The restricted availability of up-to-date and pertinent medical information impedes the chatbot's ability to furnish comprehensive and accurate responses to user inquiries.

4. Complex Reasoning and Decision-Making: Medical queries frequently necessitate intricate multi-step reasoning and integration of diverse information sources. Conventional chatbots may lack the sophistication required to perform such complex reasoning tasks effectively.

By mitigating these challenges, the proposed RAG-based medical chatbot endeavors to furnish users with precise, contextually relevant, and personalized responses to medical inquiries. Such an advancement is anticipated to significantly enhance the efficiency of healthcare delivery, elevate patient satisfaction levels, and contribute to improved healthcare outcomes.

## III. Related Work

Recent literature has seen a surge in exploration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) models in healthcare applications. A research team explored the use of RAG in medical education to boost Large Language Models (LLMs) performance. Their focus was on addressing hallucination and harmful answer generation by aligning LLMs with domain-specific tasks. They proposed a combined extractive and abstractive summarization approach for unstructured textual data using representative vectors. Leveraging RAG, the aim was to enhance content generation and responses in medical education, improving the relevance and quality of information provided by LLMs in this field.[1] One notable development is MEDQA, an open-domain multiple-choice question answering dataset tailored to medical issues. Available in English, traditional Chinese, and simplified Chinese, MEDQA underscores the requirement for extensive domain-specific knowledge, particularly in the medical field, to answer its diverse question types. The dataset encompasses both one-step and multi-hop reasoning, presenting challenges in noisy evidence retrieval and the limitations of existing open-domain question answering systems in the medical domain. The document highlights the dataset's statistics, question answering approaches, and evaluation, emphasizing the need for further research to advance open domain question answering models for medical problems.[2].

In another study, the application of large language models (LLMs) in the medical domain is explored, focusing on Chain-of-Thought (CoT) prompting for complex medical questions. It assesses the performance of open-source models like InstructGPT and Codex in zero-shot and few-shot settings. The study underscores LLMs' potential in understanding medical queries, recalling expert knowledge, and executing non-trivial reasoning. However, it also highlights the importance of cautious implementation due to potential biases and limitations in healthcare. Additionally, it discusses advancements in language models and their adaptation to prompt-based learning for new tasks without extensive fine-tuning.[3]

Med-PaLM 2, an upgraded version of their medical question answering model, building on the groundwork laid by Med-PaLM. Med-PaLM 2 integrates enhancements in base large language models (LLMs), leveraging improvements from PaLM 2, alongside domain-specific fine-tuning on medical data and innovative prompting strategies, including ensemble refinement. These enhancements significantly improved performance, with Med-PaLM 2 achieving notable accuracy of up to 86.5% on the MedQA dataset, marking a substantial advancement in medical question answering capabilities. Furthermore, the model exhibited promising outcomes across various medical question answering datasets, highlighting its effectiveness in addressing complex medical queries with high precision and efficacy.[4]

CataractBot, an LLM-powered chatbot with expert oversight, fills the gap in reliable health information accessibility. Created with a tertiary eye hospital in India, it promptly responds to cataract surgery inquiries using curated knowledge. Offering multimodal and multilingual support, it caters to diverse user needs. In a real-world study with 49 participants, CataractBot proved invaluable, providing round-the-clock assistance, saving time, and accommodating users with varying literacy levels. Trust was solidified through expert verification, underscoring the importance of credible health information. These findings could shape future endeavors in expert-mediated LLM bot design to improve health information dissemination.[5] A group of researchers introduced MultiMedQA, a benchmark addressing limitations in assessing clinical knowledge in LLMs. It combines six existing medical question answering datasets with a new dataset, HealthSearchQA. They proposed a human evaluation framework focusing on factuality, comprehension, reasoning, possible harm, and bias. Evaluating PaLM and its variant, Flan-PaLM, on MultiMedQA using various prompting strategies, Flan-PaLM achieved state-of-the-art accuracy. However, human evaluation revealed gaps, leading to the introduction of instruction prompt tuning for efficient domain alignment. Although Med-PaLM showed improvements, it still trailed behind clinicians. The study underscores the potential of LLMs in medicine, emphasizing the need for robust evaluation frameworks and method development for safe and effective clinical applications.[6] Prompt-RAG, is another novel method to enhance LLMs' performance in specialized domains without relying on vector embeddings. They compared KM and CM document embeddings, revealing distinct differences in correlation. Prompt-RAG outperformed existing models in relevance and informativeness in a QA chatbot application, despite facing challenges like content structuring and response latency. Its advancements are poised to make it a valuable tool for domains requiring RAG methods. [7]

Another study improves RAG systems by integrating fine-tuned LLMs with vector databases using LoRA and QLoRA methodologies. This approach refines models through user feedback and a Quantized Influence Measure, enhancing result selection precision. It provides insights into LLM optimization and proposes new directions for more advanced conversational AI systems.[8] Another novel RAG-LLM framework as a CDSS for safe medication prescription were developed recently. They evaluated its efficacy in identifying medication errors across medical specialties using patient case vignettes. Comparing two integration modes, autonomous and co-pilot with junior pharmacists, they found that the RAG-LLM framework, especially in co-pilot mode, significantly improved accuracy in identifying medication errors, including severe Drug-Related Problems (DRPs), highlighting its potential for enhancing medication safety in clinical practice.[9]

These studies underscored the importance of context, external knowledge, and personalized responses in enhancing the effectiveness of AI-driven healthcare interactions.

## IV. Methodology

The basic working mechanism for the retrieval augmented generation (RAG) based chatbot is depicted on the figure below:-

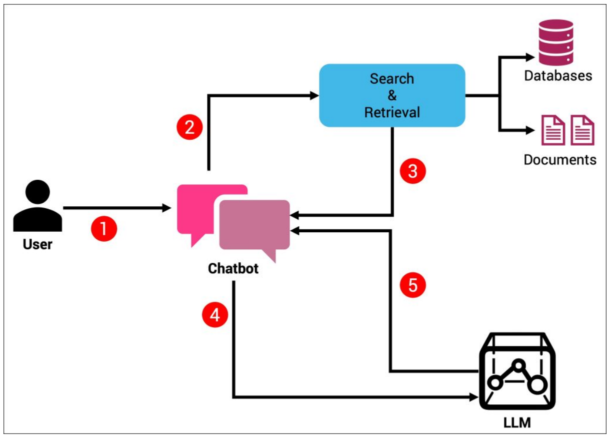
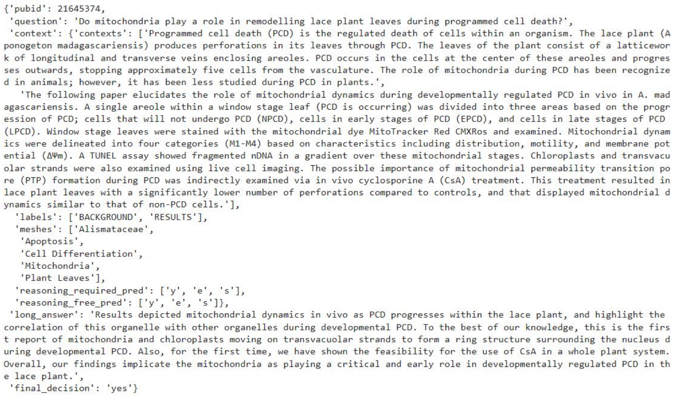


Fig. 1. The basic working mechanism for the retrieval augmented generation (RAG) based chatbot

The methodology for constructing a medical chatbot utilizing the Retrieval-Augmented Generation (RAG) model comprises several essential stages, detailed below:

### A. Data Gathering and Refinement

Firstly, it's essential to gather a wide range of relevant data from sources like medical textbooks, Q and A databases, or pertinent text repositories. Subsequently, the collected data needs to be refined to guarantee its suitability and accuracy for the specific purpose. Considering the same, we've utilized the PubMed QA repository from Hugging Face for data collection.

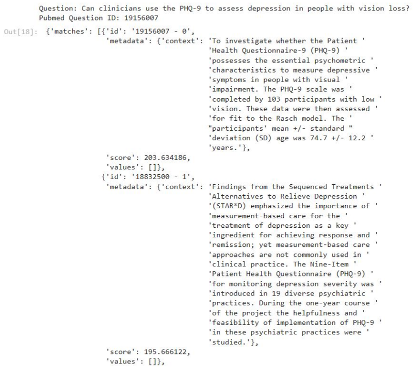


```
{'pubid': 21645374,
 'question': 'Do mitochondria play a role in remodelling lace plant leaves during programmed cell death?',
 'context': {'contexts': ['Programmed cell death (PCD) is the regulated death of cells within an organism. The lace plant (A
ponogeton madagascariensis) produces perforations in its leaves through PCD. The leaves of the plant consist of a latticewor
k of longitudinal and transverse veins enclosing areoles. PCD occurs in the cells at the center of these areoles and progres
ses outwards, stopping approximately five cells from the vasculature. The role of mitochondria during PCD has been recognize
d in animals; however, it has been less studied during PCD in plants.',
   'The following paper elucidates the role of mitochondrial dynamics during developmentally regulated PCD in vivo in A. mad
agascariensis. A single areole within a window stage leaf (PCD is occurring) was divided into three areas based on the progr
ession of PCD; cells that will not undergo PCD (NPCD), cells in early stages of PCD (EPCD), and cells in late stages of PCD
(LPCD). Window stage leaves were stained with the mitochondrial dye MitoTracker Red CMXRos and examined. Mitochondrial dynam
ics were delineated into four categories (M1-M4) based on characteristics including distribution, motility, and membrane pot
ential (ΔΨm). A TUNEL assay showed fragmented nDNA in a gradient over these mitochondrial stages. Chloroplasts and transvacu
olar strands were also examined using live cell imaging. The possible importance of mitochondrial permeability transition po
re (PTP) formation during PCD was indirectly examined via in vivo cyclosporine A (CsA) treatment. This treatment resulted in
lace plant leaves with a significantly lower number of perforations compared to controls, and that displayed mitochondrial d
ynamics similar to that of non-PCD cells.'],
  'labels': ['BACKGROUND', 'RESULTS'],
  'meshes': ['Alismataceae',
   'Apoptosis',
   'Cell Differentiation',
   'Mitochondria',
   'Plant Leaves'],
  'reasoning_required_pred': ['y', 'e', 's'],
  'reasoning_free_pred': ['y', 'e', 's']},
 'long_answer': 'Results depicted mitochondrial dynamics in vivo as PCD progresses within the lace plant, and highlight the
correlation of this organelle with other organelles during developmental PCD. To the best of our knowledge, this is the firs
t report of mitochondria and chloroplasts moving on transvacuolar strands to form a ring structure surrounding the nucleus d
uring developmental PCD. Also, for the first time, we have shown the feasibility for the use of CsA in a whole plant system.
Overall, our findings implicate the mitochondria as playing a critical and early role in developmentally regulated PCD in th
e lace plant.',
 'final_decision': 'yes'}
```

Fig. 2. Data Instance

### B. Development of a Retrieval Model

Constructing a functional chatbot necessitates the creation of a robust retrieval model. This involves data gathering, preprocessing, model selection, indexing, utilization of retrieval algorithms, validation, fine-tuning, scaling, and ongoing enhancement. The effectiveness of the retrieval model significantly influences the chatbot's ability to provide appropriate responses to user inquiries. Vector database was constructed using **just the contexts** provided within the dataset. To create a retrieval model we have considered hybrid search using vector and sparse embeddings. To generate sparse embeddings, we have considered Splade Encoder and for vector embeddings, we utilized a sentence transformer with the dimension of 768.

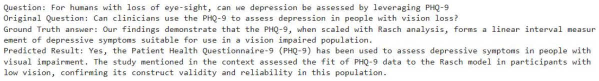


```
        Question: Can clinicians use the PHQ-9 to assess depression in people with vision loss?
        Pubmed Question ID: 19156007
Out[18]: {'matches': [{'id': '19156007 - 0',
                'metadata': {'context': 'To investigate whether the Patient '
                                        'Health Questionnaire-9 (PHQ-9) '
                                        'possesses the essential psychometric '
                                        'characteristics to measure depressive '
                                        'symptoms in people with visual '
                                        'impairment. The PHQ-9 scale was '
                                        'completed by 103 participants with low '
                                        'vision. These data were then assessed '
                                        'for fit to the Rasch model. The '
                                        "participants' mean +/- standard "
                                        'deviation (SD) age was 74.7 +/- 12.2 '
                                        'years.'},
                'score': 203.634186,
                'values': []},
               {'id': '18832500 - 1',
                'metadata': {'context': 'Findings from the Sequenced Treatments '
                                        'Alternatives to Relieve Depression '
                                        '(STAR*D) emphasized the importance of '
                                        'measurement-based care for the '
                                        'treatment of depression as a key '
                                        'ingredient for achieving response and '
                                        'remission; yet measurement-based care '
                                        'approaches are not commonly used in '
                                        'clinical practice. The Nine-Item '
                                        'Patient Health Questionnaire (PHQ-9) '
                                        'for monitoring depression severity was '
                                        'introduced in 19 diverse psychiatric '
                                        'practices. During the one-year course '
                                        'of the project the helpfulness and '
                                        'feasibility of implementation of PHQ-9 '
                                        'in these psychiatric practices were '
                                        'studied.'},
                'score': 195.666122,
                'values': []},
```

Fig. 3. Retrieval results for a given query

### C. Integration of Retrieval and Generation Models

Establishing seamless integration between the retrieval and generation components is crucial. The design must facilitate cohesive interaction between these elements to enable efficient exchange of information, thereby fostering meaningful conversations. The architecture of the RAG chatbot determines its capability to balance searching and generating content, manage content flow, accommodate user interaction, and seamlessly combine retrieval and generation processes. We are currently leveraging OpenAI's gpt-turbo-3.5 for text generation and for the retrieval database, we have considered Pinecone with a Dot Product index.



```
Question: For humans with loss of eye-sight, can we depression be assessed by leveraging PHQ-9
Original Question: Can clinicians use the PHQ-9 to assess depression in people with vision loss?
Ground Truth answer: Our findings demonstrate that the PHQ-9, when scaled with Rasch analysis, forms a linear interval measur
ement of depressive symptoms suitable for use in a vision impaired population.
Predicted Result: Yes, the Patient Health Questionnaire-9 (PHQ-9) has been used to assess depressive symptoms in people with
visual impairment. The study mentioned in the context assessed the fit of PHQ-9 data to the Rasch model in participants with
low vision, confirming its construct validity and reliability in this population.
```

Fig. 4. Ground truths and predicted responses

### D. Incorporation of Contextual Understanding

The distinguishing feature of a RAG-powered medical chatbot lies in its proficiency in comprehending and leveraging context in responses. This capability distinguishes it from conventional chatbots by providing contextually enriched information and maintaining coherent and pertinent dialogues with users. To provide better contexts for text generation, we have leveraged hybrid search for context retrieval from the database.

### *E. Testing and Maintenance*

Upon deployment of the RAG-powered medical chatbot, rigorous testing and assessment are essential to evaluate its functionality, adaptability, user interaction, and compliance with regulations. Continuous monitoring of the chatbot's performance and error logs is critical for ensuring its long-term viability. Additionally, periodic updates and fine-tuning of the medical knowledge base and RAG model are imperative to keep the chatbot accurate and up-to-date with the latest advancements in AI.

### *F. Incorporation of Advanced Features and Customizations*

Subsequent to deployment, the focus shifts to enhancing the chatbot with advanced functionalities and personalized customizations. These enhancements elevate the chatbot from a functional tool to a sophisticated and tailored solution, such as incorporating patient history-based diagnostic capabilities.

The fundamental operational mechanism of the retrieval-augmented generation (RAG) based chatbot is illustrated in the figure below.

By adhering to these systematic steps, a medical chatbot leveraging the RAG model can be developed to deliver effective, contextually aware, and accurate information to users. Overall, the methodology for leveraging RAG in medical research entails the integration of retrieval-augmented generation systems to efficiently access and generate insights from diverse healthcare data sources while addressing associated challenges and limitations.

## V. Evaluation

To evaluate a Retrieval augmented generation (RAG) pipeline, **RAGAS** was leveraged. Multiple metrics were used to evaluate the pipeline. These metrics are as follows:

- **Retrieval Metrics**
  Retrieval is the first step in a RAG pipeline, so we will focus on metrics that assess retrieval first. For that we primarily want to focus on **'context recall'** and **'context precision'** but before diving into these metrics we must understand what it is that they will be measuring.
  **Actual vs. Predicted**
  When evaluating the performance of retrieval systems we tend to compare the 'actual' (ground truth) to 'predicted' results. We define these as:
  **Actual condition** is the true label of every context in the dataset. These are positive if the context is relevant to our query or negative if the context is irrelevant to our query.
  **Predicted condition** is the predicted label determined by our retrieval system. If a context is returned it is a predicted positive, i.e $\hat{p}$. If a context is not returned it is a predicted negative, i.e $\hat{n}$.
  Given these conditions, we can say the following:
  $p\hat{p}$ is a 'true positive', meaning a relevant result has been returned.
  $n\hat{n}$ is a 'true negative', meaning an irrelevant result was not returned.
  $n\hat{p}$ is a 'false positive', meaning an irrelevant result has been returned.
  $p\hat{n}$ is a 'false negative', meaning an relevant result has not been returned.
  Let's see how these apply to our metrics in RAGAS.
  - **Context Recall**
    Context recall (or just recall) is a measure of how many of the relevant records in a dataset have been retrieved. RAGAS calculates 'Recall@K' for recall, where the '@K' represents the number of contexts returned. As the @K value is increased the recall scores will improve (as the capture size of the retrieval step increases). At it's extreme we could set @K equal to the size of the dataset to guarantee perfect recall — although this negates the point of RAG in the first place. By default, RAGAS uses a '@K' value of '5'. It is calculated as:

$$Recall@K = \frac{p\hat{p}}{p\hat{p} + n\hat{n}} = \frac{Relevant\ contexts\ retrieved}{Total\ number\ of\ relevant\ contexts}$$

  - **Context Precision**
    Context precision (or just precision) is another popular retrieval metric. We typically see both recall and precision paired together when evaluating retrieval systems.
    As with recall, the actual metric here is called Precision@K where @K represents the number of contexts returned. However, unlike recall, precision is focusing on the number of relevant results returned compared to the total results returned, whether they are relevant or not — this is equal to our chosen @K value.

$$Precision@K = \frac{p\hat{p}}{p\hat{p} + p\hat{n}} = \frac{Relevant\ contexts\ retrieved}{Total\ number\ of\ relevant\ contexts}$$

- **Generation Metrics**
  Post Retrieval, we need to evaluate generated texts for a given query. For evaluating the text generation we utilize the following metrics
  - **Faithfulness**
    The 'faithfullness' metric measures (from 0 to 1) the factual consistency of an answer when compared to the retrieved context. A score of 1 means all claims in the answer can be found in the context. A score of 0 would indicate no claims in the answer are found in the context.
    We calculate the faithfullness like so:

$$Faithfulness = \frac{Number\ of\ claims\ found\ in\ context}{Number\ of\ claims\ in\ answer}$$

  - **Answer Relevancy**
    Answer relevancy is our final metric. It focuses on the generation component and is similar to our "context precision" metric in that it measures how

much of the returned information is relevant to our original question.

We return a low answer relevancy score when:

* Answers are incomplete.

* Answers contain redundant information.

A high answer relevancy score indicates that an answer is concise and does not contain ”fluff” (ie irrelevant information).

The score is calculated by asking an LLM to generate multiple questions for a generated answer and then calculating the cosine similarity between the original question and the generated questions. Naturally, if we have a concise answer that answers a very specific question, we should find that the generated question will have a high cosine similarity to the original question.

Using these metrics, we evaluated 10 medical questions. Here are the results for the same.

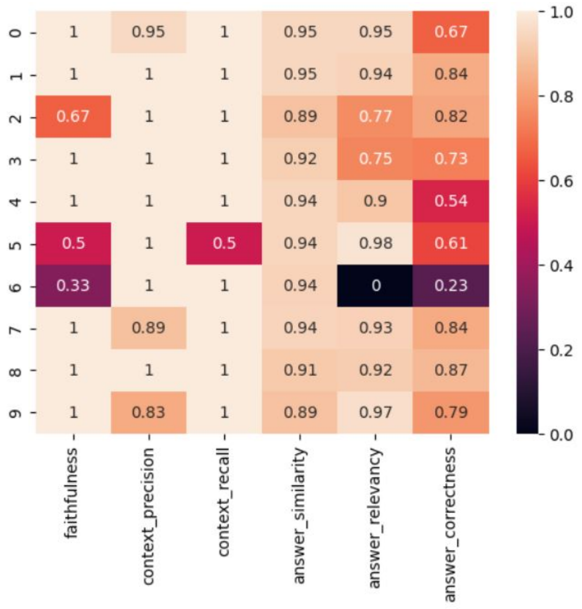


Fig. 5. Evaluation matrix for 10 medical questions

## VI. Conclusion

In conclusion, the integration of chatbots within the medical domain has emerged as a prominent trend, offering numerous benefits such as enhancing patient care, streamlining processes, and fostering better engagement between patients and specialists. While chatbots like Med-Palm, OneRemission, Ada Health, Florence, and Babylon Health facilitate communication through various channels such as phone calls, video calls, messages, and emails, it’s important to note that they are designed to complement rather than replace medical professionals. These chatbots provide round-the-clock availability, reduced waiting times, cost-effectiveness, and personalized care, making them invaluable assets in modern healthcare delivery.

However, the adoption of Large Language Models (LLMs) in medical settings presents challenges due to their tendency to disseminate inaccurate information and replicate biases. Retrieval-Augmented Generation (RAG) models offer a promising solution by leveraging specific data sources to enhance context, accuracy, and reliability. RAG models enable the development of medical chatbots equipped with contextual understanding capabilities, enabling them to provide relevant and comprehensive responses to user inquiries.

The methodology for constructing a medical chatbot using RAG involves critical steps, including data collection, refinement, creation of retrieval models, linking retrieval and generation models, implementation of contextual understanding, testing, maintenance, and incorporation of advanced features. These steps ensure the effectiveness, context awareness, and accuracy of RAG-powered medical chatbots in delivering precise information to users.

Evaluation methodologies for RAG-based medical chatbots includes assessing answer relevancy, faithfulness, context recall and context precision. By employing these evaluation methodologies, we comprehensively evaluated the performance of RAG-based medical chatbots, ensuring it’s accuracy, relevance, and contextual understanding in providing medical information and assistance.

Overall, RAG-based medical chatbots hold significant potential in advancing healthcare accessibility, efficiency, and personalized patient care.

## References

[1] S. S. Manathunga and Y. A. Illangasekara, “Retrieval augmented generation and representative vector summarization for large unstructured textual data in medical education,” 2023.

[2] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” 2020.

[3] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, “Can large language models reason about medical questions?” 2023.

[4] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, “Towards expert-level medical question answering with large language models,” 2023.

[5] P. Ramjee, B. Sachdeva, S. Golechha, S. Kulkarni, G. Fulari, K. Murali, and M. Jain, “Cataractbot: An llm-powered expert-in-the-loop chatbot for cataract patients,” 2024.

[6] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, “Large language models encode clinical knowledge,” 2022.

[7] B. Kang, J. Kim, T.-R. Yun, and C.-E. Kim, “Prompt-rag: Pioneering vector embedding-free retrieval-augmented generation in niche domains, exemplified by korean medicine,” 2024.

[8] K. Rangan and Y. Yin, “A fine-tuning enhanced rag system with quantized influence measure as ai judge,” 2024.

[9] J. C. L. Ong, L. Jin, K. Elangovan, G. Y. S. Lim, D. Y. Z. Lim, G. G. R. Sng, Y. Ke, J. Y. M. Tung, R. J. Zhong, C. M. Y. Koh, K. Z. H. Lee, X. Chen, J. K. Chng, A. Than, K. J. Goh, and D. S. W. Ting, “Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties,” 2024.