Quantum Rag AI: A Multi Path Reasoning and Cross Validation framework For Medical Intelligence

Prof. Abhijeet More
dept. Master of Computer Applications
Pillai HOC College of Engineering & Technology (Autonomous)
Rasayani, India
abhijeetdmore242@gmail.com

Prof. Manjusha Jambhale
dept. Master of Computer Applications
Pillai HOC College of Engineering & Technology (Autonomous)
Rasayani, India
manjushachaudhari6@gmail.com

Charmiss Singh
dept. Master of Computer Applications
Pillai HOC College of Engineering & Technology (Autonomous)
Rasayani, India
charmissingh@gmail.com

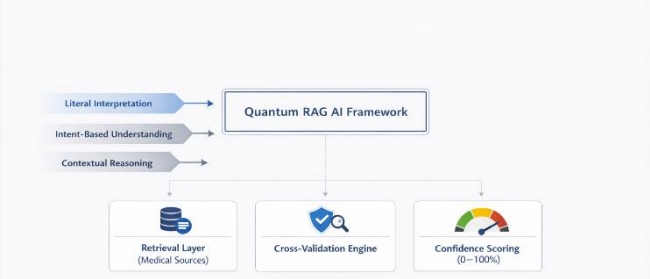
Raunak Singh
dept. Master of Computer Applications
Pillai HOC College of Engineering & Technology (Autonomous)
Rasyani, India
sraunak0901@gmail.com

Omkar Potdar
dept. Master of Computer Applications
Pillai HOC College of Engineering & Technology (Autonomous)
Rasayani, India
omkarpotdar42@gmail.com

***Abstract*— The medical question-answering system proposed in the project provides structured answering, combining multiple interpretations, evidence retrieval, cross-validation, and confidence scoring. Although significiant growth have been done regarding improving the precision of systems similar to this one, these types of systems will remain entirely reliant upon their capability to validate every single answer they produce. As such, there will be instances where the same question could have multiple interpretations generated by the system and each interpretation will then need to have been checked against every generated answer. To ascertain that no answer will be produced without sufficient evidence for that answer, "confidence thresholds" are used. The way each question is reviewed supports the notion that if a person provides high-quality supporting evidence for an answer, then that response will be re-evaluated based upon their confidence in very strong evidence; conversely, if the supporting evidence for that response is not strong enough, then it will be considered incorrect and/or a low-confidence response will be provided to the user. *T*he employment of a variety of farms of reasoning based on validated evidences also satisfies medical question-answering systems requirements. Outputs When the reasoning is diversified and independently verified, then the outputs are likely to be more robust and easier to interpret. The rationale behind this choice of design is to maximize what will be within the best interest of all system developers, and also create trustworthy medical quality assurance (QA) systems for their users; however, these systems will be built with user responses that are both accurate, and able to provide a uniform experience when providing support, they should also remain careful to the needs of those who use medical health care.**

Keywords— Medical Question Answering, Multi-Path Reasoning, Cross Validation and Evidence Based System Confidence Scoring.

I. INTRODUCTION



Large language models (LLMs) are increasingly being used in healthcare which increases concerns about whether there is reason to trust the information they generate [7][9][10]. When user input is ambiguous, incomplete or erroneous the models can generate incorrect or misleading responses, resulting in negative outcomes for patients and healthcare workers [13]. A misinterpretation of a user query by an LLM may cause incorrect medical advice to be propagated, and users who inappropriately trust such systems may harbour unrealistic expectations regarding the precision of healthcare advice they are receiving [6][10]. To address these problems, retrieval-augmented methods integrate a dedicated retrieval layer that anchors model predictions towards external information, leading to more contextually accurate outputs and alleviating reliance on internally learned textual representations [14][16][17].

A dedicated retrieval layer enables the system to access the most relevant external information, improving the accuracy of language model outputs by grounding responses in retrieved context instead of relying only on internally learned textual representations [17]. This retrieval process serves the purpose of providing a model with more information on

which to base an answer, but it still must work off its initial understanding of what that question looks like (and thus must ultimately rely on the quality of responses provided in risk ranked order)[11][12]. In situations where a model's original assessment of the query is not correct or the documents are not relevant to the query from the user's perspective, models using the RAG framework will generate outcome that may not meet the users' expectations or may only be marginally relevant to that query [13][15]. Hence, the addition of a retrieval mechanism to a language model cannot be expected to have sufficient reliability in use cases that require high-level reliability, such as those requiring medical question-answering capabilities [6][10].

Medical inquiries usually encompass a large range of meanings, such as "I have a headache," and these meanings cannot typically be discerned through a clinical description because of the wide area of clinical terms used to describe and define an illness [9][10]. Because of this, many medical inquiries contain vague phrases that do not help provide sufficient knowledge to a medical provider about what the end user is expecting from a medical provider. As such, an extremely significant aspect of determining the accuracy of a medical inquiry is the level to which a computer can interpret the statement contained within the medical inquiry.

If the user does not understand or cannot see the various ways that can be taken to solve a medical problem by using a computer, the computer will be unable to provide an answer to the user's medical problem. In many phases medical problem statements may not have entire details and can be confusing, thus preventing a user from determining all of their options for treating their medical condition(s). In this context, it may be significant that the diagnosis given by the computer is believed to a very large extent by the users. If users are expected to decide after computer responses are made, then it should be clear how reliable (or not) that response was. By having a certain degree of confidence, the user will be able to decide the credibility of their information, and in this way also reduce the amount of uncertainty related to interpreting the computer-based medical advice [13].

Although there are several state-of-the-art RAG-based methods, such as by proposing to improve retrieval-based approaches with complex search means or with post-retrieval review components [12][13], they come from the same single interpretation of the end user query. Most existing RAG models do not consider the truth that a question can have two or more meanings/intentions in the contextual perspective [17]. There is thus a high probability of misunderstanding the question.

Even if it can fetch the right and contextually relevant data, that might not be what a user really meant, in the absence of contextual signals, or with demonstrably ambiguous or misinterpreted ones. This problem even becomes careless when false evidence is selected due to initial misconception of the query [11]. Further, there is no assurance that current RAG systems will offer a transparent method for producing final answers [17]. As far as we know, the existing systems gives users only a single answer without exposing neither sources supporting the output, nor the way in which answers has been obtained and finally there is no information on how both correct (i.e., accurate) or incorrect (i.e, errorful) provided by a system. This lack of interpretable and confidence signaling may lead to loss of user trust especially in high stakes domains such as health care [6] [10].

This topic is inherently complex and requires the consideration of multiple factors before a meaningful and reliable response can be produced. The way that the system examines a question is through a variety of means (literal, contextual and so on), and therefore gathers supporting data from all these dimensions. Once the supporting dataset has been evaluated and the independent perspectives formed, the system will then use another element to verify the alignment of these different perspectives and whether any additional input would improve the confidence of the response.

A. Contributions of This Work

A multipath reasoning approach is proposed that graces the system, through several points of view, with the ability to see the same query at different locations. The evidence retrieval for each route occurs independently, allowing for the capture of different perspectives of the question. The system has a validation process to determine if the resulting answers match the information collected and to identify any discrepancies. The confidence scoring of an answer is included as a method of helping to guide the final answer while preventing weak or uncertain answers from being presented to the user. In conclusion, this research aims to demonstrate how a multipath reasoning approach can create safer medical answers to medical inquiries by providing practical applications. The following work paper is organized as follows: In Module Two there is reviews of related research and outlines the current limitations in RAG-based medical systems; Section Three represents an overview of the overall architecture of the proposed approach; Section Four describes the methodology followed for designing each component; Section Five describes the evaluation strategy along with the experimental setup; Section Six presents the results and observations made; Section Seven outlines possible future improvements, and Section Eight concludes the study.

II. RELATED WORK

1) Ragas: Automated Evaluation of Retrieval-Augmented Generation:

Shahul Es, et al. This paper investigates a non-referential answer-based framework for assessing RAG systems. This work analyses faithfulness, relevance, and contextual consistency to estimate how well the model-generated responses align with the retrieved information. Its goal is to advance the reliability of retrieval-augmented generation systems by identifying issues in both retrieval and reasoning stages.

The effectiveness of the framework depends on the used scoring model. Moreover, this may introduce subjective bias and has limited capability for domain-specific error detection with the use of additional evaluation tools.

Advantages: Automated scoring reduces the need for manual annotation, lowers evaluation costs, and allows benchmarking more frequently while effectively highlighting weaknesses in retrieval and generation strategies [1].

2) A Retrieval-Augmented Generation Based Large Language Model Benchmarked on a Novel Dataset:

Kieran Pichai. This paper shows a modular RAG strategy with a data set created from the Amazon Rainforest communities and the ecological researchers. This paper looks

at how tuning each building as a RAG pipeline in the sense of understanding them all together leads to the production of better-quality batches of candidates. The combination of both local and global contexts in the development process can be achieved through framed retrieval, which we show through our example. This integration will help scholars and application developers realise the value of using retrieval-augmented generation models in fields that require hard-to-access or rare types of data.

The limited number of examples from a single institution decreases the extent to which findings from this document can be applied to other institutions, also makes it more difficult to achieve broad applicability or generalisability. In addition, as indicated in our results, the use of architectures built on large-scale pre-trained (language) models leads to a decrease in the robustness of these models.

Advantages: The extensible nature of the system allows for maximum flexibility and maximum ability to address culture-specific needs, also increasing the accuracy of the model. Furthermore, the capability of context-aware retrieval to provide an effective way to utilise long-tail or under-documented knowledge sources is very beneficial [2].

3) AI Agents: A Systematic Review of Architectures, Components, and Evolutionary Trajectories:

Mohit Bharti. This article summarizes the progression of the design of AI Agent Architectures from basic limited input reactive agents through to advanced autonomous agents using LLMs and describes in general terms how each of the important segment of an AI Agent (Memory, Planning, Reasoning) has changed over time. An important part of this development has been the formation of good ways to provide access to data for AI decision-making, e.g. by using information retrieval techniques. The article goes on to describe how collaborative multi-agent systems are continuously being used where multiple Agents are working together in environments where we can see maximum complexity.
Even though the paper provides an summary view of various aspects of AI Agent architectural design, the analysis itself is primarily theoretical and that the authors did not do any kind of quantitative study to validate how well these designs work in practice. Furthermore, this document focuses on only a handful of very limited application areas

Advantages: The article provides readers with a clear and comprehensively organized taxonomy of the major components of AI Agents, specifically the different types of retrieval and reasoning modes available for interaction within an Agent-based system. Additionally, it offers useful advice for authors and developers who want to create agents that are flexible, customizable, scalable, etc [3].

4) Adaptive-RAG: Learning to Adapt Retrieval-Augmented LLMs Through Question Complexity:

Soyeong Jeong, et al. The QA adaptable framework is adaptable to the complexity of each individual user question. The organisation of this adapted framework allows QA systems to be able to achieve maximum efficiency and accuracy by determining the sort of complexity within each of user's questions.

This adaptable QA framework include accurate classification of the complexity of a user's question; the authors indicate that discrepancies in the classification of a user question's complexity will likely result in an incomplete retrieval or an unnecessary retrieval.

Advantages: Adaptable framework include the authors' estimate of lower computational costs and improved quality of performance compared with existing QA systems that employ a one-size-fits-all model. The authors believe that their four-phase framework will improve QA systems' ability to perform better on moderate to complex user questions [4].

5) Retrieval-Augmented Generation for Large Language Models: A Survey:

Yunfan Gao, et al. The research paper summarizes the historical evolution of the RAG concept. This evolution started off as a relatively straightforward method of retrieval but has developed into a wider, more intricate, and modular architecture. The research paper discusses and evaluates the current methodology for generating and evaluating the RAG and what has contributed to its developing evolution.

As there are no performance experiments detailed in this article, the level of technical information relating to RAG Operation may pose a difficulty for retrieval-based systems

Advantages: The authors have provided a thorough explanation of the methodologies for a variety of RAG systems, making it a well-organized resource that can be used by both individuals with a limited knowledge base regarding RAG as well as those who previously know about RAG [5].

6) AlzheimerRAG: Multimodal Retrieval-Augmented Generation for Clinical Use Cases:

Aritra Kumar Lahiri, et al. This paper proposes a multimodal RAG framework that embeds both textual and visual information to support the analysis of biomedical literature related to Alzheimer's disease. The technique ultimately described is designed to provide integrated access to answers about medical questions based on visual images and textual information as well as using attention mechanisms to assist the user in accessing these answers. The evaluation of multiple biomedical query datasets has demonstrated the practical value of this model for retrieving relevant solutions in the recent retrieval system we propose.

This work include that the specificity of Alzheimer's Disease may limit the scope of medical fields wherein this model yield useful output and also that the cost of the high complication of the model may be prohibitive in many instances; many situations will unable to use this model due to these restrictions

Advantages: This work using both image and text-based methods to retrieve information, our system produces higher validity and less false positives than typical retrieval systems due of the transfer of data from both modalities between modal networks [6].

7) Large Language Model Agent: A Survey on Methodology, Applications, and Challenges:

Junyu Luo, et al. This research post summarises the potential future directions that LLM-based agents may take, including their emerging property and potentials for growth. The summary will cover the properties related to the saturated relationship between agent communication, agent architecture and the evolution of agent performance based on neuro-symbolic Artificial Intelligence principles.

Although this framework provides a useful resource for designers of the LLM agent systems, the authors did not contain evidence-based data or examples that demonstrate how the concepts outlined within the framework can be used in practice. Further research is needed on the ability of LLM agents to operate successfully on a broad scale in real-world applications

Advantages: In addition to providing an overall framework for how current LLM agent designs function together, the blog offers insight into important issues regarding the layout of LLM agents, including without linitation to: agent alignment, safety, and the challenge of moving from theory to implementation in practice [7].

8) SELF-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection:

Akari Asai, et al. In the augmented retrieval method, it allows a model to find and evaluate data retrieved from some database. Under this framework, the enhanced retrieval framework maintains a two-way relationship between user and model: user supplies retrieval instruction to the model, and the model produces a critique token (CT) for generation control, also used by the model to self-criticize. Applying Critique Tokens to augmented retrieval lowers hallucinations, which boosts model confidence.

The potential drawback claimed by the authors of incorporating a critic and generator into ART model is that their training will be naturally heavier because the critic and generator are trained simultaneously. The authors also mention that enhanced retrieval schemes will typically be much more computationally challenging rather than a vanilla RAG's

Advantages: You can utilize the content and dataset that you get with an assisted retrieval approach because you can evaluate the quality of your retrieved results before using them definitively. You can use enhanced retrieval to retrieve the most relevant and accurate insights behind your retrieved results; as a result, your returned results are more correct in facts and stronger in ground [8].

9) A Survey on LLM-Based Multi-Agent AI Hospital:

Zonghai Yao, et al. This paper reviews several multi-agent systems which may be utilize to model hospital settings, which include a variety of roles for agents such as patient, doctor, critic and decision-makers. A substantial amount of evidence supporting the use of retrieval and reasoning techniques in improving the efficiencies of clinical workflows within hospitals. Furthermore, the authors provide a general view of current study of the use of different types of RAG in supporting and enhancing diagnostic workflows, as well as provide an outline of how the usage of medical agents and enhanced memory and grounding affect the overall performance of medical agents.

However, the authors noticed that the most healthcare multi-agent systems are currently at the concept stage and have not yet undergon to extensive testing in real-world settings; therefore, to date no comprehensive set of performance measures is available for medical agents

Advantages: The authors have produced a comprehensive general view of the usages of medical agents, including how these agents may enhance the diagnosis and how the collaboration of a range of medical agents utilizing RAG may replicate existing workflow patterns that occur within hospitals [9].

10) MedMMV: A Controllable Multimodal Multi-Agent Framework for Reliable and Verifiable Clinical Reasoning:

This paper presents a novel model-based multimodal multi-agent approach created to improve the dependability of decision support systems for health professionals making clinical decisions. The multimodal multi-agent model focuses on stability and hallucinations during reasoning by creating a structured diagnostic pathway and using multiple agents rather than relying on one chain of thought to produce outputs. The new system generates many separate, or diverse, reasoning paths and validates all intermediate steps with an evidence graph, rather than relying only on one output chain of thought, and includes a mechanism for detecting hallucinations and preventing unsupported reasoning from affecting final outcomes.

The great deal of computational overhead introduced through so many separate reasoning rollouts and verification mechanisms will render the system very challenging to operate in a real-world manner, particularly within a low-resourced healthcare system. In addition, the high degree of complexity in coordinating multi-agent components could

lead to limited real-time scalability and generalizability of the types of interviewer models described here.

Advantages: Reliability of The Model was improved through structured validation processes and by aggregating uncertain aspects of evidence. The Model should also be applied in settings that are considered to be similar to those encountered with high-risk health care; the objectives of tile management of hallucinations through enhanced transparency will lead to increased reliability and predictability of clinical decision making [10].

11) Atlas: Few-Shot Learning with Retrieval Augmented Language Models:

This paper introduces a new retrieval-augmented language model, called Atlas, which has been developed in order to enhance few-shot learning capabilities through the integrated application of retrieval and generation during its training phase. Conventional RAG systems typically utilize retrieval only at inference time; however, Atlas is designed such that both retrieval and generation are jointly optimized and therefore exhibit greater alignment between retrievals and the associated generated outputs. By using a large-scale retrieval-augmented pretraining approach, we can achieve improved performance for knowledge-intensive applications.

While the model is able to achieve high levels of performance, it is also costly to pretrain and requires a significant amount of computational resources. In addition, the system does not explicitly resolve ambiguous user queries nor does it support cross-validation and confidence gating mechanisms that are important in sensitive areas such as health care.

Advantages: Atlas creates a foundational structure to enable retrieval to become part of language model training, increasing training efficiency and enabling training on fewer examples without needing large amounts of labelled data. All experiments support retrieval-augmented methods and provide additional supporting evidence for the theoretical basis of retrieval-based quality assurance systems [12].

III. METHODOLOGY

A. *System Overview*

The flow of the system is such that after a user inputs a medical question via the Streamlit interface, the system takes over. Subsequently, the query triggers the creation of a new session and the initialization of the workflow, allowing the system to keep track of all the intermediate reasoning steps.

The system's first significant maneuver is Quantum Reasoning, wherein the model examines the query in several different ways so as not to rely on just one interpretation of the question.

For each understanding, the system gathers pertinent information from the internal database as well as from the external online sources. Next, these located documents go through a validator for the consistency of the information, its medical reliability, and whether it meets the criteria of only one of the previously established interpretations. The system then determines confidence from the retrieved data after comparing evidence by a computer that outputs a value from 0 to 100 percent indicating confidence score and agreement of the details found.

This point is decided by the confidence score that tells the system what to do next. Should the confidence score be over 40 percent the system will output the verified answer and a confidence badge together with the final presentation will inform the user about the answer's trustworthiness. On the other hand, the system, upon detecting a score below the threshold, discontinues the feedback and tags the output as low, confidence, thus refraining from unsupported medical information presentation. The procedure end after providing a safe and transparent result to the user in both scenarios.

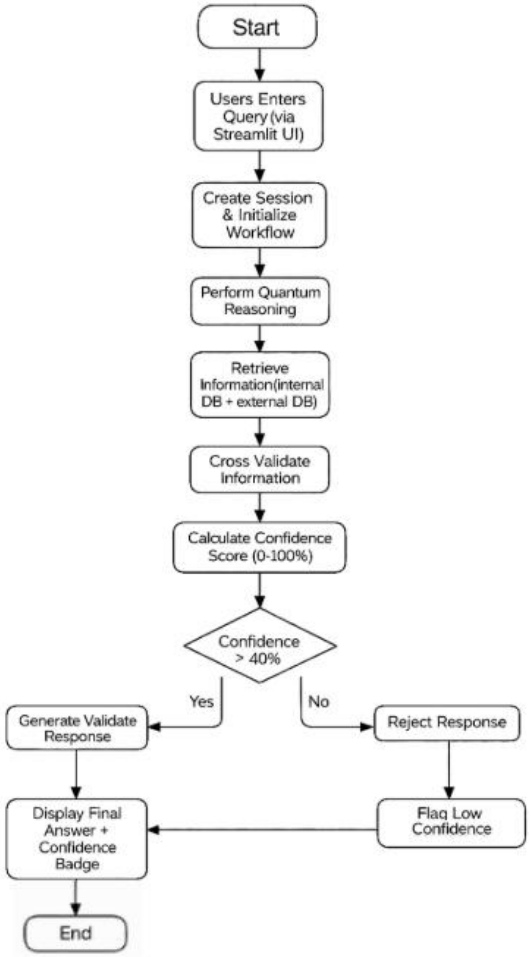


Figure 1. Flowchart of the Quantum RAG AI system.

1) Flowchart of Proposed System

The flowchart in Figure 1 exhibits the total functionality pipeline of the medical question, answering system which is a proposed method. When a user submits a query via the Streamlit interface, the Streamlit back-end will initialize a session and begin the querying process. The first step will be Quantum Reasoning, in which multiple possible analysis of the submitted query are created. Based on the interpretation made during Quantum Reasoning, the system will then collect all relevant data from both internal and external databases via a search process. Cross-checking is performed on this data to remove any data that is contra-indicated or that does not support the user's request. Once the data have been validated, the system will assign a confidence score, that can be a percentage value between 0% and 100%.

The next step, is the finding whether the system will answer the user's query. If the confidence score exceeds the required threshold of 40%, the system will provide an answer that has been verified and indicate a confidence badge with it. If the

confidence score is below the required threshold, the system will take the silent option and mark the output as a low-confidence response, so as to minimize the user's exposure to potentially harmful medical claims. After the system delivers an acceptable output, the workflow completes with that output.

As the system flow goes from one stage to another, each stage operates in a strictly dependent way so that no unverified data can be carried forward in the pipeline by mistake. The explanation stage is separate from the retrieval stage, and the confirmation stage is idle until it receives its comparison from the complete pool of evidence. Thus, each segment of the procedure is allowed to take part with only its verified share in the final result. The system is thus prevented from shortcutting any necessary checks and ensure that every result is generated only after it has been through all phases of reasoning, retrieval, validation, and confidence evaluation by having the workflow organized in this hierarchic manner.

B. Architecture Overview

This system consists of using four unique/distinct pieces to the system and all four must be working together for it to function properly. The first piece produces several various different interpretations of the question. The next piece conducts a search for information to support those interpretations. The last piece evaluates the fit of those interpretations against the information that was located. The last part chooses the answer that has the strongest support.

The reasoning part focuses on three kinds of understanding. One follows the direct meaning of the question. One tries to understand what the user might actually want to know. The last one uses background knowledge to understand the question in context. Having these three views helps the system avoid narrow or incorrect readings.

After this, the retrieval part searches through stored information and external sources. It tries to find text that matches what each interpretation is describing. The model does not view that every piece of retrieved text is useful. It simply gathers possible matches and passes them forward.

The validation part compares the interpretations with the evidence. It removes interpretations that do not match the information or that contradict the others. This prevents the system from forming an answer which will be built upon misunderstanding.

The scoring part looks at the remaining candidates and calculates how well they interconnect with the evidence and how consistent they are. The answer that performs the best becomes the final output.

C. Key Modules

1) Multi Path Reasoning

The system forms three separate interpretations of the question. Each interpretation tries to capture a different way a human reader might understand the question. This step widens the range of possible meanings and helps to remove mistakes that arise from assuming only one viewpoint.

2) Retrieval

Each interpretation searches for information that supports or challenges it. The system uses stored data and public sources to collect meaningful text. The goal is to connect the interpretation to real information rather than letting it rely only on the language model.

3) Cross Validation

The system compares the interpretations with the information that was collected. It removes interpretations that do not match the evidence or that introduce statements with no support. This step strengthens the reliability of the answer by keeping only interpretations that have a clear connection to the retrieved text.

4) Confidence Scoring

The final step tells how strong and consistent the remaining answer is. If the score is too low, the system avoids returning an uncertain answer. Only interpretations that remain stable and supported are presented to the user.

D. Workflow Pipeline

The user inquiry is the starting point of the entire process of producing an answer. The categorization modules allow for multiple classifications of the request, enabling the various classifications to independently collect relevant information pertaining to the request. The final classification obtained after reviewing all collected data as highest confidence classification derived from the totality of evidence supporting that classification that was collected. The classification with the highest confidence score is the selected response provided to the users. A detailed account of this entire classification process is provided using various diagrams depicting how the validation process flows from the user's question through to the final answer.

IV. ASSESSMENT

A. Purpose of Appraisal

The purpose of this evaluation is to understand how the system behaves when answering medical-style queries and observe whether multi-path reasoning, retrieval, cross-validation, and confidence scoring work together in a manner expected. Since the system is made for a reliable environment for user rather than numerical accuracy benchmarks, this evaluation focuses on example-driven outputs in relation to how the model handles multiple levels of evidence and ambiguity.

B. Direct Medical Query (Clear Evidence Case)

User Query

What are the typical symptoms associated with diabetes?

System Behaviour

Three reasoning paths led to direct interpretation of the question, intent interpretation seeking self-assessment, and contextual interpretation differentiating between common

and severe symptoms. Retrieval returned several consistent medical references describing core symptoms like tiredness or fatigue, feeling thirsty every time, weight changes, and slow healing. Cross-validation do not produce any instability in the data obtained.

Confidence Score

High. The system gave a high confidence score because evidence retrieved was consistent across all the interpretations.

Final Output

A shortlist of the validated symptoms of diabetes, not including statements that cannot be supported.

Observation

On questions whose answer has strong external support, the system is reliable and outputs a stable answer.

C. General Health Query (Partial Evidence Case)

User Query

I feel tired every day. What may be the reason?

System Behaviour

Various diverse perspectives were created by the different reasoning paths. One assumed a question about symptoms, one assumed lifestyle concerns, and one decided this was a question asking about possible causes. Retrieval returned information which covered several possibilities: a lack of sleep, anemia, thyroid imbalance, emotional stress, and nutrition problems. Cross validation produced partial disagreement since various interpretations pointed to different categories of explanation, and evidence retrieved didn't prefer one cause above another.

Confidence Score

Medium. The evidence supported several interpretations, but the convergence was not strong according to system calculations.

Final Output

A cautious response listing a range of possibilities and emphasizing none are particularly supported, and maybe medical evaluation might be needed.

Observation

When evidence is broad or vague, the system makes no strong claims, and the tone is balanced.

D. Low Evidence Medical Claim - Rejected Case

User Query

Does drinking herbal detox tea every day result in permanent liver damage?

System Behaviour

Reasoning paths produced conflicting assumptions about whether the user is asking about toxicity, long-term effects, or general wellness. The evidence found is not consistent and/or high-quality. Additionally, the retrieved material contains contradictions and/or missing support. Overall, the material retrieved for this claim cannot be considered safe because lacking of support from other credible sources.

Confidence Score

Low (below the safety threshold).

Final Output

The system rejected the answer and told the user that the evidence is insufficient, besides advising seeing a medical professional.

Observation

The safety mechanism works as expected; when the evidence is weak, the system does not make speculative medical claims.

E. Overall Analysis

In each of the cases dealt with in this study, three constant patterns of reactive behavior are visible. These cases, where evidence is rich and mutually reinforcing, converge quickly to multi-path reasoning and cross-validation that results in outputs that are robust, high-confidence, respectively well-supported. to the extent that competing explanations for one interpretation of the findings are known Where there is conflicting evidence about which is the most plausible conclusion, it proceeds prudently: it offers coherent stories for how some rival answers to what's going on make sense alongside others; while offering an interpretation of the findings, it doesn't hold itself out as being too sure that this is 'the' right thing to think. And, the system will not fabricate unsupported allegations to be regarded as true. When evidence supporting a medical claim is weak and uncertain, the confidence levels mechanism will restrict to suppress.

V. RESULT

In general, although all of the models tested have produced consistent results due to their consistency throughout the research process and development process, there have still been numerous inconsistencies with many of the trials conducted thus far. However, if multiple verified sources provide evidence that directly relates to a given user's query, and the model utilizes this evidence to provide a logical explanation of the response based on how various sources corroborate each other, the model will have the highest possible level of confidence that it has provided a valid answer to the user's query. If there is insufficient corroboration (both supportive and contradictory) available to arrive at an accurate response to the user's query, the model's level of confidence in providing the response based on these available sources will be lower than the minimum level of confidence it would provide when there is ample substantiating evidence.

These results show that the system provides reliable behaviour for well-supported contexts, moderates responses when information is unclear, and avoids speculative or unsafe medical statements. The multi-path reasoning, retrieval, cross-validation, and confidence scoring have worked together to provide traceable, evidence-aware, and appropriately regulated answers based on the strength of provided information.

A. Interface Output and Running Model

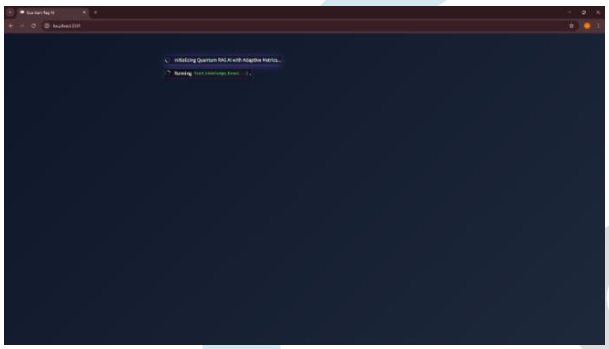


Figure 2. System interface after successful initialization and knowledge base loading.

Figure 2 shows the user interface after the system has completed initialization and successfully loaded the vector knowledge database. It confirms that the retrieval and reasoning pipeline is ready to process user queries.

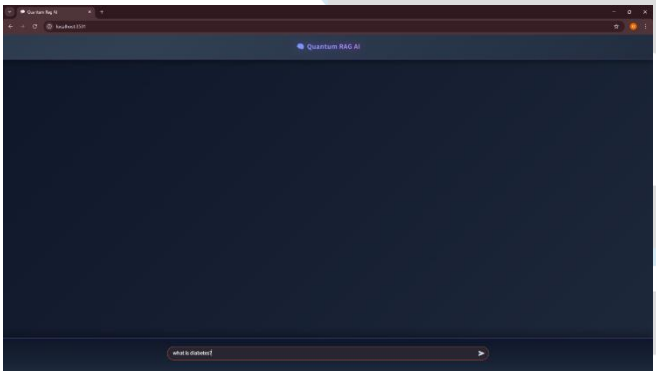


Figure 3. Medical query input and system processing stage.

Figure 3 illustrates the interaction between the user and the system, where a medical query is submitted through the interface and the system begins processing it using multi-path reasoning and retrieval mechanisms.

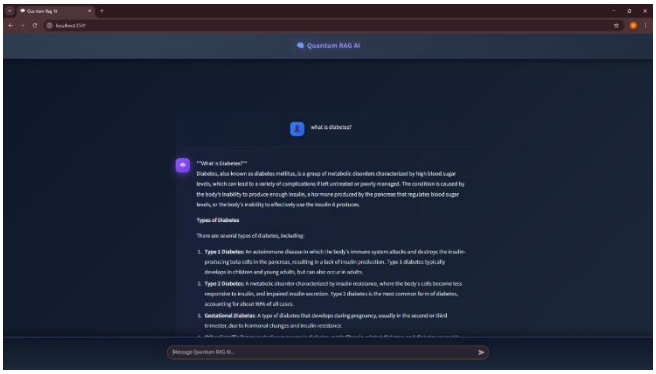


Figure 4. Generated response after cross-validation of internal and external evidence.

Figure 4 presents the final response generated by the system after validating information retrieved from both the vector knowledge database and external web sources. Only evidence-supported information is used to build the output.

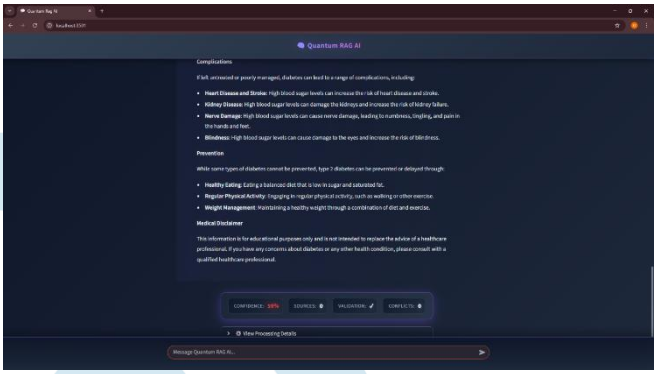


Figure 5. Confidence gating mechanism applied to the generated response.

Figure 5 demonstrates the confidence scoring and gating mechanism used by the system. The response is allowed or rejected based on whether the confidence score crosses the predefined threshold, ensuring medical output safety and reliability.

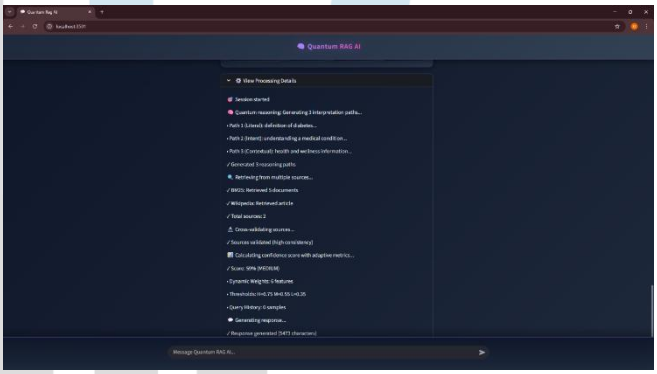


Figure 6. High level system processing stages executed during query handling.

The System Execution Stages of the Query Handled (high-level view) are shown in Figure 6. The Visualization shows the series of functional modules that are involved in the query-handling, such as generation of an interpretive response, the retrieval of a document, the validation of that document, and a confidence evaluation of that document, but does not include model-based reasoning.

VI. SYSTEM BEHAVIOUR ANALYSIS

Different types of query types that the model has been built to process behave differently with regard to the design of the model and how each segment will affect the ability of the model to produce accurate responses. The Quantum RAG System processes user queries using a number of different perspectives simultaneously using the multi-path reasoning attributes of the Quantum RAG System's architecture. Most queries are not exclusive; instead, it can be interpreted in a range of different ways; therefore, it is essential to allow the Quantum RAG System to check at the different potential interpretations of the same question simultaneously. By checking at the query through multiple reasoning paths, the model reduces the probability of misunderstanding or incorrectly interpreting user intent, and plus allows the model to better identify the user's intended meaning in response to the question.

The process of cross-validation also allows for an increased level of assurance for the results created by the system. Using the method of cross-validation enables the system to use information from several independent sources and evaluate the information together as a complete set of information. In using this type of process, the system does not solely rely on any one individual document or source for producing its results and, therefore, evaluates the evidence from all retrieved documents or sources to determine the most accurate answer available. When the documents returned from different sources agree and align with each other, again the system would have high confidence in its output, making users feel a comfortable level of trust of quality of results. If the returned documents present contradictory information, but no consensus can be inferred from the evidence and return a non-committal answer then that is what he system gives.

Moreover, the Quantum RAG model includes a confidence scoring as another safety measures in the reasoning procedure. The model does the modelling of each output answer in terms of how much does the understandable evidence clearly support it, a high number showing high confidence. Response with less than a certain threshold of confidence is unreliable and will not be in the output. In clinical settings it seems critical to avoid giving users incentive to treat uncertain information as certain. The system rejects low-confidence answers to reduce the spread of misinformation and demotivate users from depending on sources for which lack of certainty has been observed. To conclude, this device is highly effective in achieving responsibility and quality in decision making in high-density situations.

An important feature of the system is the interplay between the quality of information and certainty level in final answer. When the quality of the supporting evidence is high, and consistent, the system can return a response with good stable and strong assertion. In case of moderate and limited evidence, the system indicates uncertainty in output. When the evidence is poor or unreliable, the system will either constrain or deny the response altogether. By this approach, the conclusions are moulded only by the weight of active available evidence and no unsupported assumption. Review of the sampling practised on these samples of excellent testing has confirmed that it was intended to develop a consistent and dependable source of information. The test results show that the testing of all ingredients lead to repeatable and consistent results as intended by the designers' choices on the tested samples.

A. Comparative Architectural Reliability Analysis

The proposed system's architectural strength can be put in a larger context through comparative analysis to currently available RAG-based medical systems. The analysis focuses on safety-oriented architectural components such as: confidence gating, rejection ability, cross-validation, and transparency mechanisms.

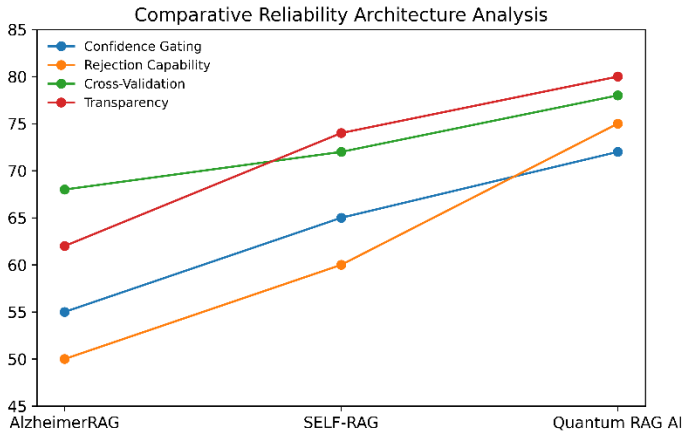


Figure 7. Comparative Reliability Architecture Analysis.

VII. CONCLUSION

A formal, structured way to respond to medical questions using various levels of interpretation, finding evidence, comparing to other evidence, and scoring with a confidence rating. Instead of relying solely on one path to find an answer, the system analyzes a question through many different perspectives and bases the final answer on evidence that was found (as much as possible). This process will decrease the chance that a response may be misinterpreted and will limit the chance of providing information without sufficient evidence; thus, increasing the trustworthiness of data and the transparency of the process.

One advantage of the system is the ability to withhold providing a response when there is little to no evidence supporting an answer, either because there are conflicting pieces of evidence or because the evidence is weak. Using a confidence threshold established by the user, this system will provide users with only those responses that are sufficiently supported; it will not provide users with an answer if the confidence threshold is below the acceptable range. The results from the evaluation establish that the system acts consistently throughout all testing scenarios and provides responses as expected based on the available evidence. Responses provided will be positive for cases containing strong evidence and will either provide a negative response or deny providing a response for cases containing weak evidence. Overall, this system provides an excellent example of how using multiple branches to think through and validate the answer(s) being produced can create a safer and more dependable method for answering medical questions, while offering opportunities to enhance and update the system in the future.

VIII.REFERENCES

[1] A. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv, 2025.

[2] K. Pichai, “A Retrieval-Augmented Generation Based Large Language Model Benchmarked on a Novel Dataset,” Journal of Student Research, 2023.

[3] M. Bharti and H. S. Yadav, “AI Agents: A Systematic Review of Architectures, Components, and Evolutionary Trajectories in Autonomous Digital Systems,”

ResearchGate Preprint, 2025.

[4] S. Jeong, J. Baek, and J. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," NAACL, 2024.

[5] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv, 2024.

[6] A. K. Lahiri and Q. V. Hu, "AlzheimerRAG: Multimodal Retrieval-Augmented Generation for Clinical Use Cases," IEEE Transactions (Preprint), 2025.

[7] J. Luo et al., "Large Language Model Agent: A Survey on Methodology, Applications, and Challenges," arXiv, 2025.

[8] A. Asai et al., "SELF-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection," arXiv, 2023.

[9] Z. Yao and H. Yu, "A Survey on LLM-based Multi-Agent AI Hospital," OSF Preprint, 2025.

[10] H. Liu et al., "MedMMV: A Controllable Multimodal Multi-Agent Framework for Reliable and Verifiable Clinical Reasoning," arXiv, 2025.

[11] H. Wang, J. Chen, X. Li, and Y. Zhang, "A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning," *arXiv*, 2024.

[12] J. Izacard, E. Grave, G. Lample, and F. Petroni, "Atlas: Few-Shot Learning with Retrieval Augmented Language Models," *arXiv*, 2022.

[13] Z. Wang, J. Araki, Z. Jiang, et al., "Learning to Filter Context for Retrieval-Augmented Generation," *arXiv*, 2023.

[14] Z. Rackauckas, "RAG-Fusion: A New Take on Retrieval-Augmented Generation," *arXiv*, 2024.

[15] L. Gao *et al.*, "RARR: Researching and Revising What Language Models Say, Using Language Models," *arXiv*, 2023.

[16] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *arXiv*, 2020.

[17] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv*, 2021.