Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 7(1) 2026, 204 – 217.

Mathematical Sciences and Informatics Journal

A Multimodal Swarm-Based Multi-Agent Retrieval-Augmented Generation (RAG) for Domain-Specific Applications

Amelia Ritahani Ismail[1*], Nurul Afiqa[1], Muhammad Izzmir Danish Zulkifli[1], Muhammad Ridhwan Irfan Nur Rashid[1]

[1]*International Islamic University Malaysia (IIUM), Gombak, 50728 Kuala Lumpur*

ARTICLE INFO

Article history:
Received 10 February 2026
Revised 17 March 2026
Accepted 24 April 2026
Online first
Published 30 April 2026

Keywords:
Multimodal AI
Medical information retrieval
Decision support
Swarm-based agents
Retrieval-augmented generation
Dental health

DOI:
10.24191/mij.v7i1.11951

ABSTRACT

The increasing complexity of medical data, encompassing text, images, and documents, requires advanced systems capable of integrating and analyzing multimodal inputs to support clinical decision-making. This study presents a novel multimodal AI system designed for medical information retrieval and decision support, using swarm-based agents and retrieval-augmented generation (RAG) to process diverse data types concurrently. The system employs collective intelligence-inspired swarm agents to distribute tasks efficiently and RAG to enhance contextual relevance. It achieved strong semantic understanding of dental health queries, with BERTScore F1 values of 0.8338 in the query with image upload scenario, though it exhibited challenges in text-image alignment with MM-Align score of 0.3103 and factual consistency of 0.4737. In query-only scenario, dental-related queries have high BERTScore F1 values of 0.9548 and mostly consistent facts of 0.950. While for non-dental queries, the system leveraged web search and give current information as result. The comparative analysis results highlight its adaptability across medical domains, despite limitations in aligning multimodal inputs and ensuring complete factual accuracy. These findings underscore the potential of swarm-based multimodal AI to improve healthcare outcomes by providing contextually relevant analyses while identifying key areas for further refinement.

[1*] Corresponding author. *E-mail address*: amelia@iium.edu.my
https://doi.org/10.24191/mij.v7i1.11951

1. INTRODUCTION

Artificial intelligence (AI) has emerged and transformed the industries sector around the world where healthcare sector also experiencing some of the most notable progress. AI-powered tools for retrieving medical information and providing decision support have become essential for assisting healthcare practitioners in diagnosing ailments, recommending treatments, and sharing knowledge with patients. Traditionally, these tools have primarily relied on organized text data, such as electronic health records or clinical guidelines to deliver responses. Nevertheless, the current medical landscape frequently demands systems capable of managing various data types such as text, images, and unstructured notes simultaneously and seamlessly. However, many of the existing systems encounter significant challenges in processing these multimodal inputs within an organized framework, particularly in question-answering contexts. This limitation restricts their potential to provide comprehensive and valuable insights for complex medical situations where visual and textual information are both critical. Moreover, present medical information systems often lack sophisticated mechanisms to dynamically retrieve and generate information customized to inquiries, frequently operating without the benefits of Retrieval-Augmented Generation (RAG) (Yang et al., 2025b). Instead, they depend on static datasets and basic methods that may not adequately capture the intricacies of medical questions (Wind et al., 2025). Furthermore, even when RAG is present, current implementations typically do not incorporate agentic RAG, which adds dynamic reasoning and action capabilities to the retrieval and generation process (Wind et al., 2025).

Current medical question-answering systems frequently encounter difficulties in efficiently integrating and reasoning across many data modalities especially visual medical images and textual data including reports and notes within a cohesive framework. This limitation restricts their ability to provide holistic and meaningful answers in complex medical inquiries that necessitate information from multiple sources and formats (Zhang et al., 2026). In addition, all the tools in the current ages often lack Retrieval-Augmented Generation (RAG) which is a method that retrieves relevant information and generates accurate as well as response that specific to the content it refers. This makes them less effective for detailed medical queries (Yang et al., 2025b). Beyond the absence of basic RAG, current systems also do not incorporate agentic RAG, which adds dynamic reasoning and action to retrieval and generation, making them less adaptable to the evolving needs of medical decision-making (Yao et al., 2023).

This research holds significant potential to change how medical information is found and how decisions are made by directly addressing the problems with current methods. By introducing a multimodal AI system with agentic RAG powered by the ReAct agent, it offers a novel approach to providing accurate, reasoning-based medical insights derived from diverse data types. This advancement could lead to a better healthcare decision which supported with evidence and improved patient outcomes through comprehensive analysis as well as greater health understanding for both doctors and patients. The system's flexible design also makes it able to adapt to various medical fields and hence be more impactful across the healthcare sector.

2. LITERATURE REVIEW

This section provides a review of related literature and all prior research that includes advanced the creation of a Retrieval-Augmented Generation (RAG) system for multimodal data processing.

2.1 Retrieval-Augmented Generation (RAG).

RAG is a hybrid framework that enhances generative language models by retrieving and integrating relevant external data to produce a meaningful and context aware responses. Yang et al. (2025a) explore the application of Retrieval-Augmented Generation (RAG) to address critical limitations of Generative

https://doi.org/10.24191/mij.v7i1.11951

Artificial Intelligence (GenAI) models in medical contexts. While GenAI systems like GPT, DALL-E and Sora show excellent impact in medical consultation, diagnosis, and education, the authors identify three fundamental challenges which include bias from pretraining data, generating misleading content and struggle to incorporate new medical research or information that specific to the patient needs.

2.1.1 RAG Framework

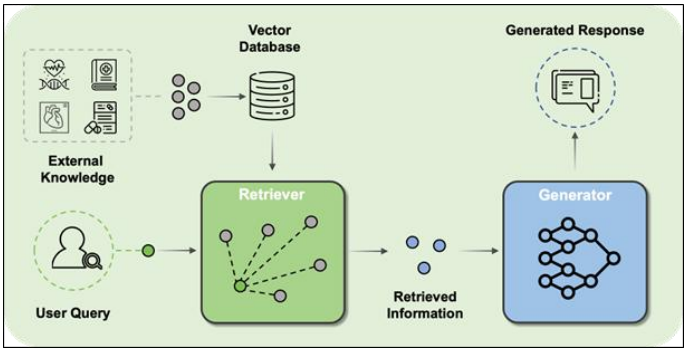

Fig. 1. RAG Architecture (Yang et al., 2025a)

Fig. 1 shows how RAG works in three steps:

(i) Indexing: External data like medical journals, guidelines, or patient records are broken into chunks and turned into vectors and stored in a database.

(ii) Retrieval: When a query is made, the system will compare it to those in the database and finds the most relevant data.

(iii) Generation: The retrieved data is fed into the GenAI model along with the query, producing a more accurate and informed response.

The integration of RAG into medical information systems marks a promising advancement toward more transparent and clinical decision that is supported with evidence. By allowing large language models (LLMs) to access medical knowledge outside the domain, RAG reduces hallucinations and enhances factual reliability in generated outputs. However, its application in healthcare remains in an early stage facing several challenges such as retrieval noise where irrelevant or low-quality evidence may influence clinical reasoning and domain shift. This causes the performance declines when models encounter unfamiliar data distributions (Neha et al., 2025). A further limitation lies in the lack of transparency of these systems, which can diminish clinician trust due to the inner process that cannot be clearly explained (Moulaei et al., 2024). Despite these barriers, RAG demonstrates substantial benefits for clinical decision-making. It improves factual consistency by grounding outputs in authoritative sources and enables greater personalization by tailoring responses to individual patient profiles (Gargari et al., 2025). Nonetheless, issues such as the lack of standardized evaluation frameworks, latency in generation, and unresolved ethical questions surrounding accountability and data provenance highlight that the deployment should be done carefully. RAG could be a game changing tool for safe and trustworthy AI-assisted medical reasoning if it is thoroughly tested and designed in a way that is easy to understand.

https://doi.org/10.24191/mij.v7i1.11951

2.2 ReAct Agent

The ReAct (Reasoning + Acting) is a framework introduced by Yao et al. (2023) which is a prompt-based method that enables a language model to alternate between thinking and doing task. It's inspired by how humans think to themselves while taking actions, like planning a task in their head and then doing it in reality. ReAct blends both by telling the model to take turns writing down thoughts like a step-by-step plan and actions like searching for information or using a tool in an interleaved manner until the task is done. In domains like medicine where accuracy and traceability are critical, ReAct's alternate reasoning and acting provides both transparency and improved the model's responses compared to approaches that rely solely on thinking or doing alone. Unlike other methods like Chain-of-Thought (CoT) where it only writes down thoughts without any actions while Action-only agents take actions but don't explain their thinking of things are done.

2.2.1 ReAct Agent Prompt

Here is how we can write an example of ReAct prompt to guide the model:

Prompt:
Answer the following questions as best you can. You have access to the following tools:
Tools:
- Google Scholar: Search academic papers. Input: research keywords.
- duckduckgo_search: A wrapper around DuckDuckGo Search. Useful for when you need to answer questions about current events. Input should be a search query.
- Semantic Scholar: Get paper summaries. Input: paper title.
Use the following format:
Use this format:
Question: [query]
Thought: [plan steps, identify needed tools]
Action: [tool name, tool input]
Observation: [tool output]
... (repeat as needed)
Thought: I have enough information.
Final Answer: [concise, evidence-based response]
Important: Never guess. Say "I don't know" if unsure.

This prompt shows how ReAct agent will guide the model to think, act, and use observations to answer a question. For our medical RAG system, we extend this template with domain-specific tools and emphasize reliable final answers to ensure clinical relevance.

https://doi.org/10.24191/mij.v7i1.11951

2.2.2 ReAct in Action: An Example

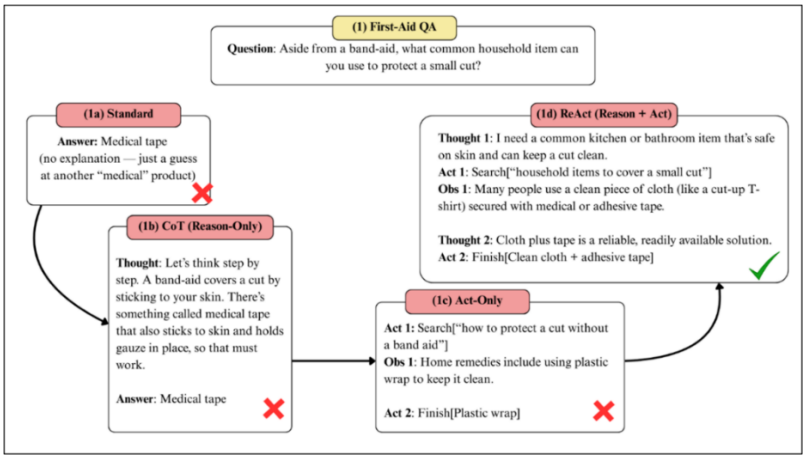

Fig. 2. ReAct Workflow Example

Fig. 2 compares different methods and shows how ReAct works step-by-step. The standard approach provided an inaccurate and useless response in this comparison by merely guessing "medical tape" without using any logic. Despite some reasoning, the CoT (Reason-Only) approach missed a more realistic household option and still came up with the incorrect answer. The Act-Only approach was based on searching and recommended "plastic wrap," which was dangerous and therefore unreliable. On the other hand, the ReAct method used both logical reasoning and factual retrieval to come up with the correct and effective solution which was "clean cloth and adhesive tape". This example highlights how ReAct's steps of decision making prevent the model from settling on an unsafe 'Band-Aid-only' answer and instead find a truly practical solution. For our medical RAG system, it shows how ReAct uses thinking and action together to find a safe, practical answer to avoid mistakes of other methods. Furthermore, ReAct agents are easy to set up, flexible and easy to troubleshoot.

2.3 Swarm-Based Multi-Agent

Swarms is an open-source platform that facilitates the development of multi-agent AI systems by enabling collaborative task-solving among specialized agents. It acts as a coordinator where it allocates specific roles, knowledge bases, and capabilities to each agent so that they can work together with ease. Unlike the traditional systems where only one agent performing tasks one after another in a fixed order, the current approach of Swarms enables agents to execute operations in parallel where it trades information with one another to achieve the same goal. This collaborative approach, inspired by biological entities like ant colonies, predisposes Swarms to complex work that requires several different kinds of expertise, such as analysing medical charts or responding to health-related queries. For this paper, Swarms' principles guide the design of Scenario 3, where agents like the Medical Data Extractor, Diagnostic Specialist, and Patient Care Coordinator collaborate to process and summarize medical documents.

https://doi.org/10.24191/mij.v7i1.11951

2.3.1 Orchestration Example

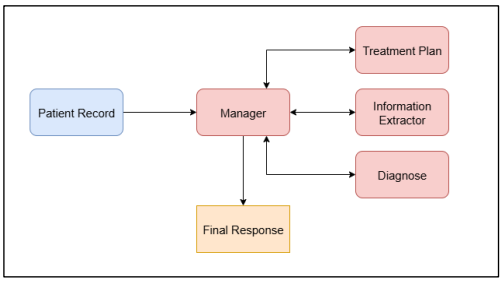

Fig. 3. Orchestration Example (Wu et al., 2023)

Fig. 3 explains a swarm intelligence model where a central Manager agent orchestrates a team of specialized AI agents to analyze a medical document. This team includes an Information Extractor for parsing raw patient data, a Diagnose Agent for identifying potential conditions and a Treatment Plan agent for recommending treatment. In this architecture, the Manager acts as the cognitive hub of the swarm where it is responsible for understanding the initial goal, tasking the appropriate agents, gathering their individual findings, and ultimately synthesizing all the outputs into a single final response. The primary advantage of such a swarm-based orchestration is that it possesses a dynamic and non-linear process. For a simple request like "Extract all medicines" or simple health record, the Manager would choose to utilize only one agent which is the Information Extractor. And while for a complex query like "Diagnose this patient and recommend a care plan," it would use all three agents, feeding the output from one into the next. This autonomy for the Manager to decide what tools to use, how much of each tool is needed, and in what order to apply them in, allows the system to formulate a "swarm" tailored to to be effective and save resources.

3. METHODOLOGY

This section outlines the method or procedure and practices used in developing the proposed system including how RAG and ReAct agents help the process. It describes how the data is processed, retrieval integration and the evaluations metrics used to measure the reliability of the results.

3.1 Swarm-Based Multi-Agents Architecture with RAG

To enable a more intelligent and context-aware response generation system, this project adopts a swarm-based multi-agent architecture that distributes tasks among specialized agents. These agents collaborate across different scenarios to process user inputs, retrieve necessary information using Retrieval-Augmented Generation (RAG), reason over the data, and generate coherent and validated outputs. The system supports diverse input types such as images, documents, and plain text queries by dynamically adapting the workflow based on the input and context requirements which are divided into three scenarios.

(i) Scenario 1 Handling image uploads with query: In this scenario, RAG is implemented to retrieve additional context from database or doing web search if the context is not available in the database. It enhances the input-handling routine to support retrieval augmented generation (RAG) with an agentic controller that assist to retrieves latent context from a local knowledge base.

(ii) Scenario 2 Processing plain text queries only: In this scenario, RAG and React Agent is utilized for information retrieval from databases or doing web searches. This scenario happens when the user input a textual query only, without any accompanying document or image upload. To

https://doi.org/10.24191/mij.v7i1.11951

achieve this, the design leverages two complementary mechanisms: i) Retrieval-Augmented Generation (RAG) for context enrichment from a local database and ii) ReAct Agent for the use of the web searches, whenever the system's internal information is incomplete."

(iii) Scenario 3 Processing query with document: Swarm technique is implemented to handle complex task which is perform comprehensive clinical analysis of a patient's health record. It establishes a team of specialized AI agents, including a data extractor, a diagnostic specialist, a treatment planner, and a consultant. A "Patient Care Coordinator" acts as the orchestrator, managing the workflow by reviewing the collected information at each step and deciding which specialist to call next. The agents collaborate by adding their findings to a shared "workspace," iteratively building a comprehensive analysis until the orchestrator determines the case is complete and generates a final summary.

Each scenario demonstrates how various agents interact with each other. The system accesses databases if the query relates to self-data or performs web searches if it does not, before working with the Gemini model for advanced reasoning and response generation.

3.1.1 Scenario 1: Image Upload with Query and Scenario 2: Query Only

Fig. 4 shows Multi-Agents Workflow that supports two scenarios which are image-based input and textual queries. This architecture leverages a ReAct Agent, Retrieval Agent, Web Search, Chroma DB, Gemini and an Answer Validator Agent to determine the best pathway for producing a reliable answer.

In Scenario 1 (Image Upload with Query - purple path), the user uploads a dental image (e.g., showing a cavity) along with a query. Gemini, the multimodal model, analyzes both the image and text to detect abnormalities. The input is then sent to the Retrieval Agent, which queries Chroma DB for relevant medical knowledge. Gemini integrates the retrieved context with its initial analysis to generate an informed response. This output is validated by the Answer Validator Agent using ROUGE for text and MM-Align for multimodal alignment. The validated answer, enriched with contextual and visual information, is then delivered to the user.

In Scenario 2 (Text-Based Query - blue path), for text-only queries (e.g., "What causes cavities?"), the ReAct Agent directs the process. If the query matches entries in Chroma DB, the Retrieval Agent extracts relevant information, which Gemini uses to generate a domain-specific response. The Answer Validator Agent then evaluates the result using ROUGE and MM-Align before presenting it to the user. If the query does not match any internal data (yellow path), the ReAct Agent triggers a Web Search. Relevant external content is retrieved, passed to Gemini for response generation, and validated before being shown to the user. This ensures that both self-data and external knowledge can support the query.

https://doi.org/10.24191/mij.v7i1.11951

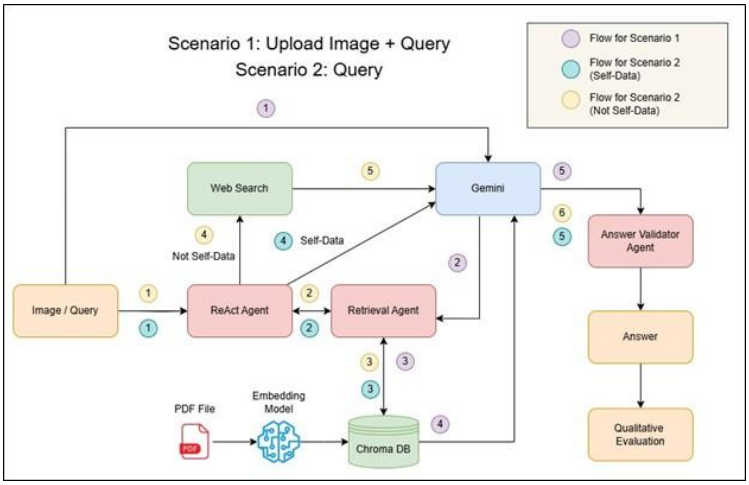

Fig. 4. Multi-Agents workflow

3.1.2 Scenario 3: Document Upload with Query

Fig. 5 defines several AI agents with specific roles: a Data Extractor to pull key information, a Diagnostic Specialist to form a diagnosis, a Treatment Planner to suggest care plans, and a Specialist Consultant to provide deeper insights. A central "manager" agent, the Patient Care Coordinator, acts as the orchestrator, directing the entire process based on the specific goals of the analysis. The workflow operates in a dynamic loop where the Patient Care Coordinator reviews all information gathered in a shared "workspace." Based on this context, it makes an intelligent decision about the next logical step. This is not a fixed pipeline; the orchestrator has the flexibility to decide which tool to use first and how many tools are needed for the task. For a simple query, it might only deploy the Data Extractor. For a complex analysis, it will sequentially call upon multiple agents. Each specialist adds its findings back to the workspace, progressively building a complete analysis until the coordinator determines the task is complete, issues a "FINISH" command, and synthesizes all reports into a final summary.

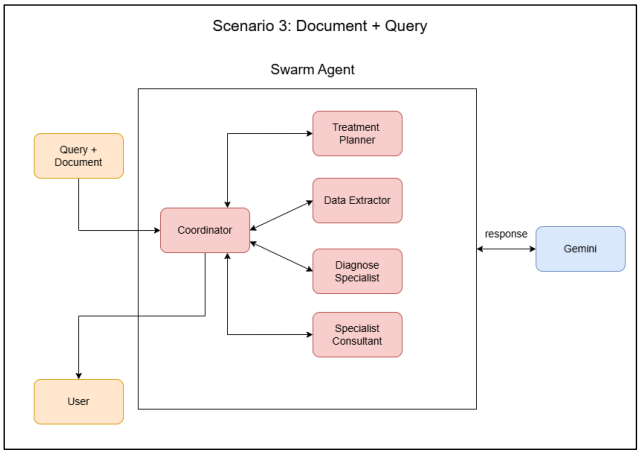

Fig. 5. Swarm-Agent workflow

https://doi.org/10.24191/mij.v7i1.11951

3.2 Dataset

The Medical Question Answering Dataset (MedQuAD) (Afroz, 2023) was utilized during model development to provide a reliable foundation of domain-specific knowledge, which is the medical. Carefully selected question-answer pairs from twelve reliable sources run by the U.S. National Institutes of Health (NIH), such as Cancer.gov and the Genetic and Rare Diseases Information Resource (GARD), are included in this dataset. Every entry contains information that has been medically verified and is intended to support clinical and patient-centered research. The MedQuAD dataset was ingested into ChromaDB, the system's vector-based knowledge repository, for this investigation. During response generation, this structure enables the Gemini model to retrieve, cross-reference, and analyse pertinent medical data.By grounding outputs in verified manner the model enhances the accuracy, credibility, and contextual reliability of its medical answers thereby supporting trustworthy and reproducible AI-assisted medical dialogue.

3.3 Evaluation Metrics

To assess the performance and quality of responses generated by the RAG and Gemini system, we employed a comprehensive set of evaluation metrics. These metrics evaluate various dimensions of the system's outputs, including text similarity, semantic equivalence, multimodal alignment and factual correctness. The following metrics were used:

(i) ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Using n-gram matching, ROUGE measures the overlap between the response generated by the system and the reference response. It calculates precision, recall, and F1-score for unigrams (ROUGE-1), bigrams (ROUGE-2), and the longest common subsequence (ROUGE-L). The F1-score for ROUGE-1, for instance, is defined as:

$$ROUGE - 1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

ROUGE provides a reliable measure of surface-level text similarity but may not fully capture semantic equivalence or meaning.

(ii) MM-Align (Multimodal Alignment Score)

MM-Align evaluates the alignment between text and visual content using cosine similarity of CLIP embeddings. It assesses how well the response corresponds to the visual input, the alignment of the reference text with the visual input and the similarity between the generated answer and the reference. The cosine similarity between two embeddings a and b is calculated as:

$$cosine\ similarity = \frac{a \cdot b}{||a||\ ||b||} \tag{2}$$

where a ·b is the dot product of the vectors, and ||a|| ||b|| are their magnitudes. MM-Align is essential to ensure coherence between text and images.

(iii) BERTScore (Bidirectional Encoder Representations from Transformers Score)

BERTScore uses contextutal embeddings from BERT to calculate the semantic similarity. It captures deeper linguistic meaning by computing cosine similarity between token embeddings. The F1-score for BERTScore is given by:

https://doi.org/10.24191/mij.v7i1.11951

$$BERTSCORE\ F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

This metric is particularly valuable for assessing the quality of rephrased or paraphrased responses produced by Gemini, where semantic equivalence matters more than exact word overlaps.

(iv) Factual Consistency

Factual consistency evaluates whether the generated response aligns factually with the retrieved context hence addressing the risk of hallucinations in generative models. This metric ensures that the system provides reliable and truthful outputs

$$Factual\ Consistency = \frac{Number\ of\ Consistent\ Response}{Total\ Responses} \quad (4)$$

4. RESULTS AND ANALYSIS

This section depicts the output of each case along with the performance analysis in terms of each of the evaluation metrics. The description is based on how retrieval and multimodal features affect the quality and contextually accurate responses generated.

4.1 Result and Analysis

Table 1 is an example of query that is evaluated using reference answer and generated answer.

Table 1. Reference and Generated Answer

	Reference Answer	Generated Answer
What are cavities?	Cavities, or tooth decay, are permanent damage to the tooth's hard surface, forming tiny holes due to bacteria and poor oral hygiene.	Cavities, also known as dental caries or tooth decay, are permanently damaged areas in the hard surfaces of teeth. They develop into holes and are caused by bacteria, frequent snacking, sugary drinks, and poor oral hygiene.
What is periodontal disease?	Periodontal disease, also known as gum disease, is a chronic bacterial infection affecting the gums and supporting bone of teeth. It starts with gingivitis and can progress to periodontitis.	Periodontal disease, also known as gum disease, is a chronic bacterial infection affecting the gums and supporting bone of teeth. It starts with gingivitis and can progress to periodontitis.

4.1.1 Scenario 1: Query + Image Upload

Table 2. Evaluation Metrics for Scenario 1 (Query + Image Upload)

Metric	Query 1: "What happens to my teeth?"	Query 2: "Is this a cavity?"
BERTScore F1	0.8217	0.8338
MM-Align (Answer-Image)	0.3042	0.3103
MM-Align (Ref-Image)	0.2876	0.2876
MM-Align (Answer-Ref)	0.9148	0.8269
Factual Consistency	0.6842 (Inconsistent)	0.4737 (Inconsistent)

BERTScore F1 values (0.8217 and 0.8338) indicate strong semantic similarity between generated answers and the references which means that the system captures meaning well despite image complexity. However, low MM-Align scores shows that there is a poor alignment between text and image, possibly due to image quality or lack of specific visual references. High Answer-Reference MM-Align scores (0.9148

https://doi.org/10.24191/mij.v7i1.11951

and 0.8269) confirm textual alignment with references. Factual Consistency ratios below 0.7 indicate inconsistencies which might be cause by missing key details despite semantic similarity.

4.1.2 Scenario 2: Query Only

In this scenario, users submit a query of text only about dental conditions. The results evaluated by each metrics are summarized in Table 3.

Table 3. Evaluation Metrics for Scenario 2 (Query Only)

Metric	Query 1: "What are cavities?"	Query 2: "What is periodontal disease?"
ROUGE-1	0.3373	0.4186
ROUGE-2	0.0741	0.2381
ROUGE-L	0.2651	0.3256
BERTScore F1	0.9786	0.9548
Factual Consistency	0.8571 (Consistent)	0.9501 (Consistent)

ROUGE-1 score, 0.3373 and 0.4545 show moderate word overlap, but lower ROUGE-2 scores, take 0.0741 suggest rephrasing or synonym use reduces exact matches. High BERTScore F1 values in both query which are 0.8808 & 0.9786 confirm strong semantic performance. Factual consistency is mostly consistent (0.75 & 0.9501).

While for queries outside the dental domain, the system employs a web search tool. Since no predefined references exist, quantitative metrics are not applicable. Instead, qualitative evaluation was conducted:

Query: What is the latest treatment for cancer?
Output: Recent cancer treatment advances include CAR T-cell therapy, showing decades-long remission in some leukemia patients, and dostarlimab, receiving Breakthrough Therapy Designation from the FDA for rectal cancer after showing tumor disappearance in some patients. Additionally, an mRNA-based cancer vaccine shows potential in stimulating long-lasting immune responses against pancreatic cancer.

4.1.3 Scenario 3: Medical record documents using a swarm-based multi-agent system

Query: Do not ask suggestions or ask any improvement for the clarity because this is a finalized medical record. Extract key medical information, diagnose, plan treatment, provide specialist insight, and coordinate a care plan based only on this document.

Output: This summary outlines your recent diagnosis and next steps in your care.

Diagnosis: You have been diagnosed with type 2 diabetes. This means your body doesn't use insulin properly, leading to high blood sugar levels. This diagnosis is based on your blood tests showing high fasting blood glucose and HbA1c levels, as well as your symptoms.

Symptoms: Your symptoms, which include persistent fatigue, increased thirst and urination, unexplained weight loss, and blurred vision, are consistent with type 2 diabetes.

Risk Factors: Several factors increase your risk for type 2 diabetes, including your family history (father and maternal grandmother had type 2 diabetes), your past gestational diabetes, and your current lifestyle (limited exercise and frequent consumption of processed foods and sugary drinks). You also have high blood pressure and high cholesterol, which are often associated with diabetes.

https://doi.org/10.24191/mij.v7i1.11951

Next Steps: You will need to see an endocrinologist (a hormone specialist) and a registered dietitian (nutrition counselor) for further evaluation and management of your diabetes. These specialists will help you develop a personalized plan to manage your blood sugar levels. Further evaluation is also planned to assess for possible nerve damage (diabetic neuropathy) related to your diabetes. You will receive comprehensive diabetes education to help you understand your condition and how to manage it effectively.

Medications: You are currently taking Lisinopril for high blood pressure and Atorvastatin for high cholesterol. Continue taking these medications as prescribed.

Lifestyle Changes: Managing your diabetes will involve making some lifestyle changes. These changes may include improving your diet by reducing processed foods and sugary drinks, increasing your physical activity (aim for more than 20-minute walks more frequently), and monitoring your blood sugar levels regularly. Specific recommendations will be provided by your endocrinologist and dietitian.

Follow-up Care: You will need regular follow-up appointments with your primary care physician and the specialists to monitor your blood sugar levels, blood pressure, and cholesterol, and to adjust your treatment plan as needed. The frequency of these appointments will be determined by your healthcare providers.

The system shows strong capability in extracting and summarizing core medical information from documents, producing concise and patient-oriented outcomes meeting query requirements. However, Scenario 3 has serious evaluation problems since the absence of quantitative metrics preventing objective measurement of strength of performance, as well as qualitative tests like LLM-based fact-checks, cannot guarantee medical correctness. Therefore, expert verification remains indispensable to make sure the clinical correctness of the outputs generated. Hence, while secondary LLM verification indicated proper summarization, this method alone is insufficient for the system where the error occurs could have serious consequences. The absence of ground-truth medical validation remains a critical limitation. Therefore, domain expert involvement is very crucial for reliable assessment.

5. CONCLUSION

This study presented a multimodal AI system for medical information retrieval and clinical decision support where it involves integrating swarm-based agents with agentic RAG through the ReAct and Swarm-based multi-modal RAG framework. Three scenarios are used in the evaluation process which include queries with image uploads, queries alone, and queries with document uploads. The system demonstrated improved distributed processing and dynamic reasoning capabilities. However, challenges remain in achieving seamless multimodal integration and ensuring factual reliability. Overall, the findings highlight the value of combining retrieval, reasoning, and multimodal processing to enhance adaptability and utility in healthcare AI systems. Future developments will focus on refining swarm-agent coordination which will further improve processing efficiency and accuracy in complex tasks. Additionally, linking the model to real-time medical databases will ensure outputs remain aligned with current evidence and treatment guidelines.

6. ACKNOWLEDGEMENT/FUNDING

The authors would like to acknowledge the support of Kulliyyah of Information Communication and Technology (KICT), International Islamic University Malaysia (IIUM) for providing the facilities on this research.

https://doi.org/10.24191/mij.v7i1.11951

7. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

8. AUTHORS' CONTRIBUTIONS

Conceptualization, A.R.I., M.I.D and M.R.I.; Methodology, A.R.I., M.I.D and M.R.I.; Formal analysis, A.R.I., M.I.D and M.R.I.; Writing—original draft preparation, A.R.I., N.A.; writing—review and editing, A.R.I., N.A. All authors have read and agreed to the published version of the manuscript.

9. DECLARATION OF GENERATIVE AI IN THE WRITING PROCESS

The authors declare that generative artificial intelligence (AI) tools may have been used solely to support language refinement during the preparation of this manuscript. Such tools were used only to improve grammar, sentence structure, clarity, and overall readability. All aspects of the research, including the study design, data collection, data extraction, analysis, interpretation of findings, and final academic decisions, were conducted, reviewed, and approved by the authors. The authors take full responsibility for the accuracy, integrity, originality, and scholarly content of this manuscript.

10. DATA AVAILABILITY/SUPPLEMENTARY MATERIALS

The data and/or supplementary materials supporting the findings of this study are available from the corresponding author upon reasonable request, where applicable.

11. ETHICS STATEMENT

The authors declare that this study did not involve human participants or animal subjects. Where applicable, all procedures were conducted in accordance with relevant institutional guidelines, safety requirements, and ethical standards.

REFERENCES

Afroz. (2024). *MedQuAD: Medical question-answer dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/pythonafroz/medquad-medical-question-answer-for-ai-research

Gargari, O. K., & Habibi, G. (2025). Enhancing medical AI with retrieval-augmented generation: A mini narrative review. *Digital Health, 11*, 20552076251337177. https://doi.org/10.1177/20552076251337177

Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., & Afrash, M. R. (2024). Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics, 188*, 105474. https://doi.org/10.1016/j.ijmedinf.2024.105474

Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI, 6*(9), 226. https://doi.org/10.3390/ai6090226

Wind, S., Sopa, J., Truhn, D., Lotfinia, M., Nguyen, T.-T., Bressem, K., Adams, L., Rusu, M., Köstler, H., Wellein, G., Maier, A., & Tayebi Arasteh, S. (2025). Multi-step retrieval and reasoning improves

radiology question answering with large language models. *npj Digital Medicine, 8*, Article 790. https://doi.org/10.1038/s41746-025-02250-5

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv*. https://doi.org/10.48550/arXiv.2308.08155

Yang, H., Chen, J., & Zhou, Y. (2025a). Enhancing clinical text generation through retrieval-augmented architectures. *Journal of Medical Systems, 49*(2), 55–66.

Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W., & Liu, N. (2025b). Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems, 2*, Article 2. https://doi.org/10.1038/s44401-024-00004-1

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

Zhang, R., Chen, Y., Yue, W., Zhang, Y., Li, X., Feng, S., Yuan, F., & Luo, M. (2026). Multimodal artificial intelligence in medicine: A task-oriented framework for clinical translation. *Frontiers in Medicine, 12*, Article 1736272. https://doi.org/10.3389/fmed.2025.1736272

https://doi.org/10.24191/mij.v7i1.11951