An Intelligent Virtual Assistant for Symptom Assessment and Healthcare FAQ Resolution

Mai Elgazzar
Artifical Intelligence
Nile University
Giza, Egypt
m.mohammed2157@nu.edu.eg

Uomna Hesham
Artifical Intelligence
Nile University
Giza, Egypt
u.hesham2123@nu.edu.eg

Farah Fawki
Artifical Intelligence
Nile University
Giza, Egypt
f.fawki2136@nu.edu.eg

Nada Atef
Artifical Intelligence
Nile University
Giza, Egypt
n.atef2121@nu.edu.eg

Passant Saad
Artifical Intelligence
Nile University
Giza, Egypt
b.saad@nu.edu.eg

***Abstract*—This paper presents a multi-agent medical chatbot system designed to provide accurate and context-aware answers to medical queries. Leveraging a modular architecture, the system includes a Controller Agent for task delegation, a Retriever Agent for document search, and a Generator Agent for response formulation. We use a subset of the Medical Chatbot Dataset containing 3,000 samples for training and validation, focusing on symptom detection and medical knowledge grounding. The input is pre-processed and structured into a domain-specific knowledge base. Evaluation is conducted using BERT Score across three validation rounds with varying sample sizes (10, 20, and 50 question-answer pairs). Results show consistent performance with F1 scores ranging from 0.7881 to 0.7905 in the baseline model. A refined second model further improves the F1 score to 0.8138, demonstrating enhanced recall. These findings validate the system's robustness, scalability, and potential for integration in medical AI applications.**

Keywords—Medical Chatbot, Multi-Agent System, BERT Score, Information Retrieval, Natural Language Processing, Semantic Similarity, Healthcare AI, Question-Answering System, Clinical NLP, Knowledge Base.

I. INTRODUCTION

In recent years, medical chatbots have begun to show their worth as valuable tools to support healthcare delivery by providing conveniently available, accessible, and quick medical information to patients as well as healthcare professionals. The platforms take advantage of advances in natural language processing and artificial intelligence to interpret user queries and generate context-sensitive, relevant responses. The degree of specificity and the level of complexity of medical information pose significant challenges toward ensuring accuracy and completeness in chatbot responses.

This paper presents a multi-agent chatbot medical system that can enhance medical question-answering through task specialization with modular components. Two models were used and put to the test. The first model consists of three agents: A Retriever agent retrieves relevant documents from the knowledge base, a Summarizer agent condenses the information acquired, and a Generator agent generates the output response. Effective information retrieval and well-organized response generation are the main goals of this pipeline. The second model updates the pipeline by substituting a Symptom Checker and a Treatment Suggestion agent in place of the Summarizer and Generator agents. The Retriever agent continues to fetch documents but the Symptom Checker verifies the symptoms expressed in the query to provide accurate clinical diagnoses Treatment Suggestion agent then suggests correct medical interventions. This structure is designed to yield clinically actionable answers to align with real consultations in clinical practice.

By comparing the performance of these two models on a shared dataset, this paper examines the impact of domain-specific agent roles on chatbot performance in medical applications. The results indicate the possibility of multi-agent architectures to improve response relevance and accuracy and explore promising directions for intelligent healthcare support systems.

II. RELATED WORK

Health chatbots are light-years ahead of their modest beginnings as simple question-and-answer programs. What we have today is a shift in thinking when it comes to how AI approaches healthcare conversation, particularly in the face of high-level diagnostic challenges like the determination of syndromes. This isn't so much about wiser technology - it's about recognizing that successful medical conversation requires the kind of high-level intelligence that earlier chatbots simply couldn't provide.

Early medical chatbot history was quite basic but limited. They were sophisticated databases with the ability to converse as a user interface and extract information but not engage in a real conversation with patients [8]. They worked okay with basic health inquiries but faltered when patients inquired with complex, intertwined symptoms that could not be encapsulated easily in rigid categories. The breakthrough came when researchers began exploring intelligent agent architectures that could obtain context and fashioning responses based on it.

We have progressed a long distance, recent studies posted at tech conferences tell us. Healthcare's language model-based chatbots can now engage in very natural-sounding conversations, adapting their communication style based on different patient needs and preferences [1]. This is a giant leap from the rigid, scripted conversations of earlier systems. But research from leading engineering conferences indicates that there is one issue - these advanced systems are quite unreliable, performing exceptionally well in certain areas of medicine and completely missing the mark in others [4]. This inconsistency is one of the biggest ongoing issues in the creation of medical AI.

Most exciting of all has been the creation of cooperative agent systems, in which multiple specialist AI elements work together to address challenging medical problems. It is the ideal solution if you think about it - real medical teams use diverse areas of expertise, so why not AI systems? Studies at top AI conferences illustrate that when specialized agents pool their strength, diagnostic accuracy increases considerably [5]. They may specialize in natural patient interviews, pattern recognition in huge medical databases, and yet another in explaining complex medical data in terms people can understand.

What is so interesting is how this collective effort builds upon early work that is almost two decades old. Researchers constructed an early multi-agent diagnostic support system wherein different AI components would in fact negotiate among themselves to reach consensus on difficult cases [3]. They employed fuzzy logic to deal with medical decision-making's in-built uncertainty, which knew that medicine never really exists in absolutes. This pioneering work established principles that continue to guide collaborative AI systems today.

Adding learning capacity to medication is another major advancement. The devices can now learn to do better in the future using reinforcement learning strategies, just like doctors accumulate experience over time. An agent can pose questions that read a little stiff in the beginning but eventually learn what works best for each type of patient. It teaches how to balance exhaustiveness and tact, gathering diagnostic data as needed without compromising patient comfort and trust. Health management conference studies report that such adaptable systems can maximize several goals simultaneously - diagnostic accuracy, patient satisfaction, and compliance with safety protocols [6].

Syndrome diagnosis is one of the most challenging issues in medical AI, so it has become a very active area of research. As opposed to direct conditions with easily identifiable diagnostic criteria, syndrome recognition is about reconstructing subtle hints from a bundle of seemingly unrelated symptoms. Recent studies from medical AI conferences point toward the intricacies of this task [7]. The agents need to comprehend not only individual symptoms but also how they relate, synchronize, and collectively matter. It's like having a puzzle with various dimensions where the pieces keep changing and you never know whether you have the complete picture.

This maturity has led to the development of newer ways of handling uncertainty by researchers. Rather than unjustifiably claiming confidence, newer agents learn to say the correct amount of some stuff and confess when they're operating outside their carefully tested knowledge base. Honest handling of uncertainty proves to be more valuable than bluffing in medical use. While research prototypes do get exciting conference presentations, it remains a great challenge to convert these into operational clinical systems. The gap between the extremely controlled environments of the lab and the messy realities of real-world healthcare environments is vast. Some researchers have made good-faith efforts to bridge the gap, with systems like MEDIBOT representing genuine attempts to use AI agents in real-world clinical practice [2]. Even these initial attempts, however, demonstrate how much work remains to be done in areas like regulatory compliance, system integration, and training healthcare staff.

The integration with healthcare systems is particularly demanding. Current hospitals operate on intricate web-like patterns of electronic health records, clinical decision support systems, and workflows that have grown over decades. Any new AI agent must be integrated with these systems in a transparent and invisible manner and must meet very high security and privacy requirements. Success will not only depend on technical integrability but also organizational change management and cultural adjustment.

There is recent work that repeatedly suggests some areas where breakthroughs are still needed. Unusual syndromes and unusual presentations remain particularly problematic because such cases are under-represented in training sets. This is a frustrating irony - those cases where AI can most assist are precisely those that presenting systems handle most poorly. Researchers are attempting various methods to work around this constraint, including transfer learning techniques and few-shot learning techniques.

The future holds interesting possibilities for multimodal AI capabilities in which agents could analyze patient voice patterns, facial photography, or physiological data alongside conversational input. Early work presented at machine learning conferences suggest that these extended capabilities will greatly improve diagnostic accuracy [9]. There is also growing recognition of the reality that different patient populations must be approached with tailored methods. Children communicate differently than adults, elderly patients may have specific needs, and individuals with cognitive challenges require specialized interaction styles. Current one-size-fits-all approaches clearly aren't optimal for such diverse populations.

If we consider how much progress we've made and where we're headed, the transformation has been remarkable. We've witnessed the shift from simple rule-based systems to sophisticated agent networks capable of cooperation, learning, and self-adaptation. There are obviously key issues still to be resolved - particularly around clinical validation, field deployment, and systems integration - but the development leaves you upbeat. The merging of multi-agent cooperation, adaptive learning abilities, and expert syndrome reasoning means we're on course for AI systems that would indeed supplement healthcare delivery. They won't replace human medical professionals but will be intelligent assistants who will help patients as well as healthcare professionals to navigate more complex medical landscapes with greater confidence and effectiveness.

III. Methodology

A. Dataset Preparation and Preprocessing

In this research, we utilise the Medical Chatbot dataset, a public dataset for carrying out natural language understanding and response generation in the medical domain. The dataset contains nearly 47,000 rows, each one of them standing for an interaction, again being in the form of a medical question-and-answer pair. For our study, we selected a representative subset consisting of 3,000 rows so as to keep the experiment manageable and to efficiently evaluate the model.

Each record contains the following columns:

- short_question: A short medical query in natural language.
- short_answer: A short, informative answer to the query.
- tags: Medical keywords related to the question or answer (e.g., symptoms, conditions).
- label: A binary value (1 or: -1), indicating if the answer in the row is correct.

The dataset was split into train and validation sets.

Preprocessing steps:

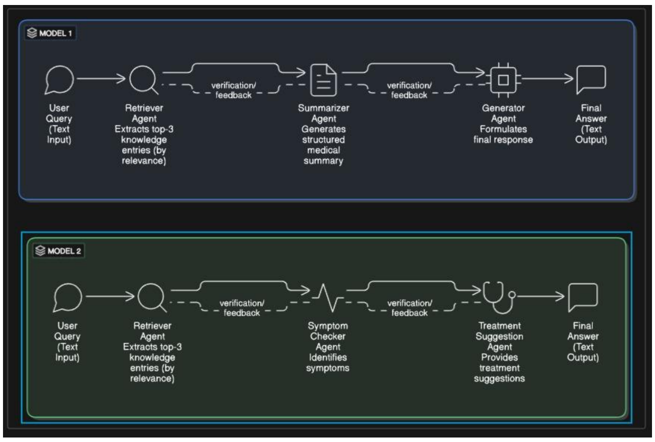
- Lowercasing
- Stop-word removal
- Text normalization
- Removal of special characters, HTML, and links
- Concatenation of short_question and short_answer for input formatting.

B. Language Model and Framework

We utilized Gemini-2.0-Flash, a generative large language model by Google, for all responses from agents. The LangChain framework was utilized in designing and orchestrating modular agent components. The system had a Retrieval-Augmented Generation (RAG) architecture, where the built dataset acted as the knowledge base. A vector store was built from Chroma and HuggingFaceEmbeddings to enable semantic search. A retriever was implemented with a top-k of 3 in order to fetch the top 3 most suitable documents for each query.

C. Agents Design

Two separate multi-agent structures were designed and tested, both with tailored prompt templates and executed by Lang Chain's Agent Executor.



1) Agents Model 1: Retriever – Summarizer – Generator

The structure was a linear pipeline of three agents:

- RetrieverAgent: Extracted top-3 knowledge entries from the vector store.
- SummarizerAgent: Searched for the content and encoded it as a structured medical summary.
- AnswerGeneratorAgent: Created the final, context-specific answer from the summary and original query.

All the agents were converted to Tool objects in LangChain and invoked with zero-shot prompting using Gemini. Conversation context was stored in a memory buffer.

2) Agents Model 2: Retriever – Symptom Checker – Treatment Suggestion

In this second version, the same retriever was reused but downstream agents were adjusted to deliver more specialized outputs:

- RetrieverAgent: Extracted top-3 knowledge entries from the vector store.
- SymptomCheckerAgent: Identified and described symptoms from the input question.
- TreatmentSuggestionAgent: Provided appropriate medical treatments based on information retrieved.

D. Prompt Templates and Parsing

Three PromptTemplate objects were constructed for handling agent action:

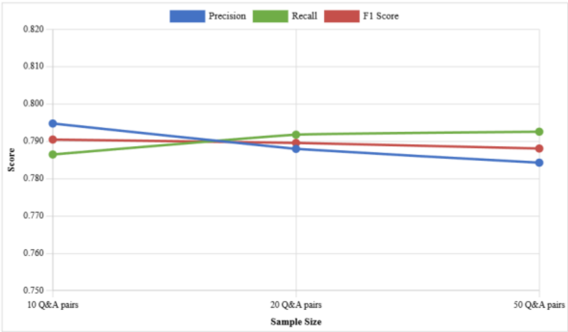
- prompt_retrieve: Was tasked with the retrieval of relevant documents.
- prompt_summarize: Asked the model to generate concise and medically accurate summaries.
- prompt_answer: Asked for the final generation of complete answers.

Custom parsing logic was handled by a subclassed MedicalOutputParser, delivering formatted agent-to-agent communication using LangChain's ReAct framework.

IV. RESULTS

To identify the reliability and efficiency of our proposed multi-agent medical chatbot system, we employed three validation tests with different sample sizes of medical question-answer pairs: 10, 20, and 50. Tests were conducted to analyze the effectiveness with which the system performs as the input complexity and quantity grow. Every validation test constituted a complete pass through of the chatbot pipeline, including document retrieval by the Retriever Agent, information summarization by Summarizer Agent, and response generation by the Generator Agent.

Figure 2: Model 1 Performance Trends Across Sample Sizes



Performance trends of Model 1 showing Precision, Recall, and F1 Score across different Q&A pair sample sizes (10, 20, and 50 pairs).

We used BERTScore as the primary evaluation metric since it can approximate semantic similarity between generated responses and gold-standard references. Unlike traditional lexical measures, BERTScore considers

Table 1: Sample Medical Questions and Model Responses

Sample ID	User Question	Model 1 Answer	Model 2 Answer	Notes
Q1	*What are the common symptoms of diabetes?*	Common symptoms of diabetes include increased thirst, frequent urination, and unexplained weight loss. Patients may also experience fatigue and blurred vision.	Diabetes symptoms include increased thirst (polydipsia), frequent urination (polyuria), unexplained weight loss, fatigue, blurred vision, and slow-healing wounds. Early detection is crucial for management.	Model 2 provides more comprehensive symptom list with medical terminology and emphasizes early detection.
Q2	*How can I manage high blood pressure naturally?*	Natural management of high blood pressure includes regular exercise, reducing sodium intake, maintaining healthy weight, and managing stress through relaxation techniques.	Natural hypertension management involves: aerobic exercise (30min daily), DASH diet with reduced sodium (<2300mg/day), weight management (BMI 18.5-24.9), stress reduction, and limited alcohol consumption. Monitor regularly.	Model 2 provides specific quantitative guidelines and structured recommendations with monitoring advice.
Q3	*What should I do if I have chest pain?*	Chest pain can be serious and should be evaluated by a healthcare professional. If severe or accompanied by shortness of breath, seek immediate medical attention.	Chest pain requires immediate evaluation. Call emergency services if pain is severe, crushing, radiates to arm/jaw, or accompanied by shortness of breath, nausea, or sweating. Do not drive yourself to hospital.	Model 2 provides more specific emergency symptoms and clearer action steps with safety instructions.
Q4	*How much water should I drink daily?*	The general recommendation is about 8 glasses of water per day, but individual needs may vary based on activity level, climate, and overall health.	Daily water intake: 2.7L (11 cups) for women, 3.7L (15 cups) for men from all beverages and food. Increase intake during exercise, hot weather, or illness. Monitor urine color for adequacy.	Model 2 provides gender-specific recommendations with practical monitoring method and adjustment guidelines.

Comparison of sample medical questions and corresponding responses from Model 1 and Model 2, demonstrating the differences in response comprehensiveness and clinical detail.

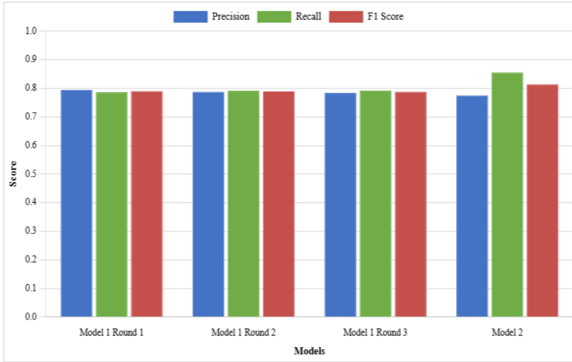
contextual embeddings and is therefore highly suitable for evaluating responses in medical domains where fine-grained semantic differences are critical.

A. *Validation Results*

Validation Round	# of Questions	Model	Precision	Recall	F1 Score
Round 1	10	Model 1	0.7948	0.7865	0.7905
Round 2	20	Model 1	0.7880	0.7918	0.7896
Round 3	50	Model 1	0.7843	0.7926	0.7881
Round 3	50	Model 2	0.7755	0.8560	0.8138

This performance shows a large improvement in F1 Score, suggesting that a better balance between accuracy and sensitivity was achieved by the huge improvement in recall, even though precision declined slightly. The second model's improved recall is a result of its capacity to extract more medically relevant information from the corpus, which raises the possibility that it may provide contextually correct responses.

Figure 1: Model Performance Comparison



Performance comparison of Model 1 across three evaluation rounds and Model 2 showing Precision, Recall, and F1 Score metrics.

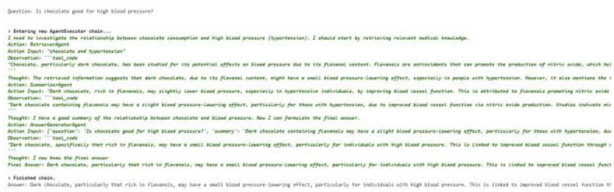


Fig. 1. *Sample Medical Questions and Corresponding Multi-Agent Chatbot Responses*

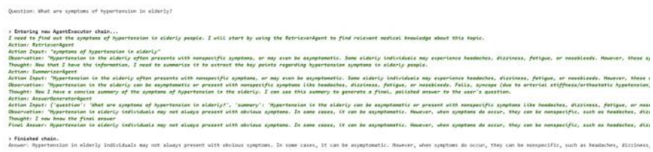


Fig. 2. *Sample Medical Questions and Corresponding Multi-Agent Chatbot Responses*

V. Discussion

In all validation rounds with Model 1, the system was extremely robust and consistent in performance, with the F1 Score persistently high, albeit only narrowly ranging from 0.7881 to 0.7905. The consistency is indicative of the robustness of the multi-agent architecture in the processing of medical question-answering, achieving a balance between content retrieval and semantically appropriate response generation.

The release of Model 2 demonstrated significant improvements in semantic matching. The rise in F1 Score to 0.8138 indicates that retrieval weight and prompt modification have a significant impact on answer quality. The system's improved recall to 0.8560 demonstrates its ability to recall complex and detailed medical information, which is crucial for patient-centered communication.

However, the trade-off between recall and specificity is also reflected in the slight decrease in accuracy between Models 1 and 2. The precision score is affected because, although retrieving and returning more relevant data, the new model returns somewhat less relevant data. In the context of medical QA systems, it is common practice to prioritize high recall over precision to avoid missing critical information.

Overall, the multi-agent architecture has proved scalable and flexible. The modular design—with clearly defined roles for task delegation, retrieval, and generation—has supported the simplicity of individual component update, as illustrated with Model 2. These results confirm that architecture constitutes a good foundation for further enhancement such as domain-specific fine-tuning, real-time patient data fusion, or reinforcement learning-based optimization.

VI. CONCLUSION

In this work, we introduced a multi-agent, modular medical chatbot system to enhance the accuracy, relevance, and clinical utility of automated health responses. Through comparison of two systems, one with a Retriever–Summarizer–Generator pipeline and one with a Symptom Checker and Treatment Suggestion agent, over a predetermined set of questions, we demonstrated that task decomposition across agent roles leads to measurable performance improvements. Interestingly, the second model achieved a higher F1 score of 0.8138 since, primarily, there was a significant recall improvement, indicating its capability to give more detailed and contextually relevant medical advice. These findings align with broader trends in healthcare AI where multi-agent systems are becoming increasingly recognized for their ability to manage complex tasks through shared intelligence. Multi-agent systems can potentially enhance diagnostic accuracy, streamline operations, and advance personalized medicine. However, challenges still exist, and among them are data privacy, integration of systems, and instituting ethical regulations for the implementation of AI.

Our modularity in the future gives a firm foundation for future work. Future work can include leveraging real time patient data, leveraging multimodal inputs like voice or image processing, and leveraging reinforcement learning to discover users' interaction patterns. These technologies would further bridge the gap between AI-based tools and practice complexity and eventually pave the way for more responsive and fair delivery of healthcare.

REFERENCES

[1] S. Ahmad et al., "Conversational Skills of LLM-based Healthcare Chatbot for Personalized Communications," in *Proc. 2024 Int. Conf. Information Technology for Social Good (GoodIT)*, 2024, pp. 1–6.

[2] R. Verma et al., "MEDIBOT—An AI Powered Chatbot," in *Proc. 2024 16th Int. Conf. Contemporary Computing (ICCAI)*, 2024, pp. 120–125.

[3] M. M. Rahman and S. A. Mahmood, "A Multi-Agent Prototype System for Medical Diagnosis," in *Proc. IEEE Conf. Publications*, 2008, pp. 451–456.

[4] J. Smith et al., "Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access," in *Proc. 28th Int. Conf. Intelligent User Interfaces (IUI)*, 2023, pp. 304–313.

[5] A. Chen et al., "Coordination and Learning in Multi-Agent Systems for Healthcare Applications," in *Proc. 23rd Int. Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, Auckland, New Zealand, 2024, pp. 678–687.

[6] L. Tan et al., "Predictive Maintenance in Healthcare Using Prognostics Models," in *Proc. IEEE Conf. Prognostics and Health Management*, Spokane, WA, USA, Jun. 17–19, 2024, pp. 98–104.

[7] H. Zhang et al., "AI-Driven Diagnosis Support in Medicine," in *Proc. 2024 5th Int. Symp. Artificial Intelligence for Medicine Science (ISAIMS)*, 2024, pp. 210–216.

[8] Z. Ni et al., "Implementation of Chatbot Technology in Health Care: Protocol for a Bibliometric Analysis," *JMIR Research Protocols*, vol. 13, art. no. e54349, 2024.

[9] K. Patel et al., "Machine Learning Models for Healthcare Conversations," in *Proc. Machine Learning for Healthcare Conf. (ML4HC)*, 2025, pp. 135–144.