International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

A Safety-Bounded Multi-Agent Medical Assistant: Retrieval-Augmented Generation, Medical Vision, and Human Validation

Dr. K. Himabindu[1], Faruk Ahmed[2], Pankaj Kumar Bara[3], Krishna Kumar Yadav[4], Soham Parmar[5]
[1]Associate Professor, Dept. of CSE PIET, Parul University Vadodara, Gujarat, India
[2,3,4,5]Department of CSE PIET, Parul University Vadodara, Gujarat, India

Abstract- Healthcare chatbots are gradually moving beyond simple question-and-answer systems by bringing together large language models, information retrieval, multimodal analysis, and human oversight. This paper explores the open-source Multi-Agent Medical Assistant as a safety-focused tool for providing medical information. The system includes separate components for input screening, conversation, retrieval-augmented generation (RAG), web-based evidence retrieval, medical image processing, response generation, and human validation. Rather than attempting to provide autonomous diagnoses, the system emphasizes evidence-based responses, transparent communication of uncertainty, and human review.

Keywords- Healthcare Chatbots, Large Language Models (LLMs), Information Retrieval, Multimodal Analysis, Human Oversight, Multi-Agent Medical Assistant, Retrieval-Augmented Generation (RAG), Medical Image Processing, Autonomous Diagnosis, Evidence-Based Responses, Communication of Uncertainty, Human review.

I. INTRODUCTION

Healthcare information systems are increasingly being used to support patients at different stages of their healthcare journey, including before, during, and after clinical consultations. Patients may use chatbots to understand their symptoms, ask questions about medications, find relevant medical guidelines, upload medical images, or prepare questions for discussions with doctors. Earlier research on healthcare chatbots mainly focused on symptom classification. In these systems, user input was processed using natural language processing (NLP), compared with a predefined set of symptoms, and then classified using techniques such as decision trees, support vector machines, random forests, and neural networks [1]–[4]. Although these approaches can be helpful for basic triage and patient education, they provide limited support for the wider range of healthcare information and assistance that patients may require today.

Large language models (LLMs) have demonstrated their ability to generate clinically relevant responses, but they can also produce information that is outdated, overly confident, unsupported by evidence, biased, or incomplete. Therefore, evaluating clinical LLMs involves more than simply measuring their performance on multiple-choice questions. Human evaluation is important for examining factors such as factual accuracy, potential harm, bias, and consistency with established medical and scientific knowledge, as highlighted by Singhal et al. [5]. Similarly, Nori et al. assessed the safety and calibration of GPT-4 using a range of challenging medical and anatomy-related problem sets [6].

These limitations suggest that medical assistants need workflows that can guide and control language-model responses instead of depending only on more powerful models. Retrieval-augmented generation (RAG) is one approach that can help with this by allowing a model to use information retrieved from external sources in addition to the knowledge already stored in its parameters. Lewis et al. described RAG as a framework that combines a model's parametric knowledge with information retrieved from a non-parametric memory [7]. Later research and surveys have explored more modular RAG architectures, where different tasks such as document parsing, chunking, retrieval, reranking, and response generation are handled by separate components [8]. This approach is especially useful in healthcare, where responses need to be based on reliable and relevant evidence.

The Multi-Agent Medical Assistant is an open-source project that provides a practical setting for exploring this approach.

Instead of depending on a single chatbot to handle every task, the system distributes different responsibilities across coordinated components. For example, one component can screen the user's input, while others retrieve relevant information from available sources, including PubMed. The system also includes separate components for medical-image classification and segmentation. Before the final response is delivered to the user, it can pass through output guardrails that help identify and limit potentially unsafe or unsupported responses.

In this paper, the project is examined as a research-oriented design study rather than as a clinically validated diagnostic system. The study focuses on how an agent-based medical assistant can be designed to make its responses more traceable, handle uncertain cases through appropriate escalation, support human review of high-risk outputs, and assess safety-related behavior. Based on this focus, the main contributions of this work are as follows:

- A synthesized architecture for agent-based medical assistance that organizes different tasks into coordinated components.
- A repository-grounded RAG workflow that connects generated responses with relevant supporting evidence.
- A safety framework that addresses task routing, uncertainty handling, output guardrails, and human review.
- An evaluation protocol covering information retrieval, response faithfulness, medical-image processing, task routing, and escalation behavior.
- A discussion of system limitations and academic-integrity practices for documenting, analyzing, and evaluating open-source medical AI systems.

II. RELATED WORK

The relevant literature covers several areas, including healthcare chatbots, clinical evaluation of large language models (LLMs), retrieval-augmented generation (RAG), agent-based workflows, and medical image analysis. These areas are discussed separately because the proposed assistant is designed as a coordinated system rather than as a single classifier or standalone chatbot. It brings together multiple components, with each component handling a specific task within the overall workflow.

Healthcare Chatbots and Symptom Classifiers

The healthcare chatbot studies reviewed in this work can be broadly divided into three research directions. The first focuses on improving access to basic medical guidance. Mahajan et al., Patil et al., and Mendapara et al. presented systems that process users' descriptions of symptoms and use NLP and machine-learning techniques to suggest possible diseases or provide general precautionary advice [1]–[3]. These studies are primarily motivated by challenges such as limited access to healthcare professionals, overcrowded hospitals, and the need for quick access to basic medical information. The second research direction focuses on algorithmic approaches to disease prediction. Athulya et al. explored decision-tree-based classification for predicting diseases from reported symptoms, while Singh et al. compared commonly used supervised learning methods using symptom-based datasets [4], [9].

These studies demonstrate that structured symptom information can be useful for basic disease classification. However, their performance depends heavily on the coverage, quality, and representativeness of the datasets used for training and evaluation. In addition, these approaches generally treat a medical consultation as a single-step classification task. In real-world settings, medical questions often involve several additional steps, such as clarifying symptoms, retrieving relevant information, communicating uncertainty, and deciding when professional medical consultation may be necessary. The third research direction looks at a broader role for healthcare chatbots.

Wanjari et al. explored chatbot-based support using neural networks and ChatterBot, while Sai Roshan et al. examined chatbot design in relation to data preprocessing and electronic health records. Pillai and Nalamwar also investigated the use of chatbots for chronic-disease management and responses based on machine-reading-comprehension approaches [10]–[12]. This line of research moves healthcare chatbots beyond simple symptom classification toward more document-grounded medical assistance. However, many of these systems still rely on static knowledge sources and offer limited support for multimodal inputs, up-to-date medical evidence, and validation by healthcare professionals.

LLMs in Clinical Question Answering

Recent studies on large language models (LLMs) have shown their potential for answering clinical questions, while also highlighting several risks that need to be considered. Singhal et

al. introduced MultiMedQA, a benchmark for evaluating medical question-answering systems based on factors such as factual accuracy, agreement with medical knowledge, potential harm, and bias [5].

Their work highlights the importance of human-centered evaluation of clinical LLMs rather than relying only on automated performance metrics. Nori et al. reported strong performance from GPT-4 on challenging medical problems, while also emphasizing the need to examine calibration, memorization, and safety risks when these models are used in high-stakes medical settings [6]. These findings point toward the need for a safety-bounded assistant that can distinguish between questions that can be answered directly and cases that require additional information, external evidence, or referral to a healthcare professional.

Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) is particularly relevant to healthcare because medical questions often require information that is current, reliable, and easy to verify. Lewis et al. showed that retrieval can improve knowledge-intensive language generation by allowing a model to access information from an external document index instead of relying entirely on the knowledge stored in its parameters [7].

Gao et al. further described how RAG has developed from basic retrieval-based approaches into more advanced and modular pipelines that incorporate techniques such as query transformation, hybrid retrieval, reranking, and systematic evaluation [8]. For healthcare applications, a modular RAG pipeline is useful because its individual stages can be inspected and verified. For example, reviewers can examine document parsing, chunk metadata, retrieval scores, reranked evidence, generated claims, and final citations to better understand how a response was produced.

Agentic Workflows

Agentic language-model systems extend the capabilities of language models by enabling them to reason, use tools, and interact with external data sources and environments. ReAct showed that combining reasoning with actions can improve tool use while making the process more interpretable in question-answering and decision-making tasks [13]. This idea can also be applied to medical assistants, although healthcare applications require stronger safety controls. Instead of allowing a medical agent to perform unrestricted actions, the system can follow a defined workflow in which each node has a specific responsibility, clearly defined inputs and outputs, and an appropriate mechanism for escalation when a case requires further review.

Medical Imaging

Medical image analysis introduces additional challenges because errors in image interpretation can directly affect how users understand their health information. The repository includes workflows for chest X-ray classification and skin-lesion segmentation, while brain-tumor analysis is also included as part of the broader system architecture. These tasks are related to existing research on brain-tumor detection [14]–[16], chest radiography for COVID-19 [17]–[21], and skin-lesion segmentation [22], [23].

However, interpreting medical images requires greater caution than providing general health information. False-positive or false-negative classifications, as well as inaccurate segmentation, may influence how users interpret their results and make decisions about their health. Therefore, image-analysis outputs should be treated as preliminary results that require appropriate validation rather than being presented as definitive diagnoses.

III. PROJECT OVERVIEW

The open-source project combines a browser-based interface and FastAPI backend with document ingestion, Qdrant-based retrieval, web search, speech processing, and computer-vision agents. The architecture separates the interface, evidence, reasoning, and control layers, allowing different parts of the system to handle specific responsibilities. User text or images are first screened and then routed to the appropriate specialist component. The selected component can retrieve relevant evidence, process the input, and pass its results through confidence checks and validation before a response is generated. This separation helps distinguish the technical capabilities of the system from the level of clinical trust that can reasonably be placed in its outputs.

IV. DESIGN PRINCIPLES

The architecture is evaluated based on four main design principles:

Bounded Medical Claims

The assistant should avoid giving definitive diagnoses, prescribing treatments, or suggesting that users delay emergency medical care. Instead, it should focus on summarizing available evidence, explaining general medical concepts, and recommending professional medical attention when symptoms are severe, persistent, unclear, or potentially urgent. This principle also has an important engineering implication: the final response layer should be able to reduce the level of certainty in an answer, even when an upstream model generates highly confident language.

Evidence Provenance

Medical responses should maintain a clear link between individual claims and the sources supporting them. In a RAG workflow, source metadata can include details such as the document name, section, page number when available, chunk identifier, and references to related images or tables. Maintaining this provenance makes it easier for users and reviewers to distinguish information supported by retrieved evidence from conclusions that are generated by the model without direct supporting evidence.

Modular Accountability

Each component should be evaluated both independently and as part of the complete workflow. For example, the router can be assessed based on its agent-selection accuracy, while the RAG component can be evaluated using retrieval quality and citation-faithfulness measures. Similarly, the vision models can be assessed using appropriate classification or segmentation metrics. The final response layer can then be evaluated using safety-related criteria. This modular approach makes it easier to identify where errors occur and provides a clearer basis for evaluating the system than treating it as a single monolithic chatbot.

Human Escalation

The assistant should defer to human review when the available evidence is limited, the potential risk is high, or medical-image interpretation is involved. Human-in-the-loop validation is therefore treated as an important safety mechanism rather than simply an optional interface feature. The repository's validation workflow follows this principle by passing outputs from the image-analysis agents through a validation prompt before considering the results complete.

V. SYSTEM ARCHITECTURE

The proposed architecture uses a directed workflow rather than a conventional linear chatbot pipeline. A graph-based controller maintains the conversation state, selects the appropriate next node, and determines whether the task is complete, requires clarification, or should be escalated for human validation. This structure allows different components to handle specific tasks while maintaining control over the overall interaction. Fig. 1 presents an overview of the proposed design.

Fig. 1. High-level architecture of the safety-bounded Multi-Agent Medical Assistant, showing guarded input analysis, task routing, specialized agents, validation, and final response guardrails.

1. Input Analysis and Guardrails

The workflow begins by analyzing the user's input. Text is first checked by the input guardrails before being passed to the appropriate agent. For uploaded images, the API verifies the file type and size before sending the image to the relevant analysis component. Inputs that are considered unsafe or irrelevant are stopped at this stage and handled with a controlled response.

2. Agent Router

The router considers the user's input, recent conversation history, image availability, and image type when selecting the appropriate agent. Depending on the request, it can route the task to the conversation, RAG, web-search, brain-tumor analysis, chest X-ray analysis, or skin-lesion analysis component. The routing decision can also be recorded along with its confidence level, providing useful information for later review and auditing.

The routing function is expressed as:

$$a* = \arg\max_{a \in A} P(a \mid x, h, m), \quad \ldots(1)$$

Here, $x$denotes the current user input, $h$represents the recent conversation history, $m$represents the available modality information, and Adenotes the set of available agents. When the routing confidence falls below a defined threshold, the system follows a safer default path rather than making an uncertain routing decision.

3. Specialized Agents

The conversation agent handles greetings, clarification requests, and basic health education. When the local knowledge base does not provide sufficient information, the web-processing agent can retrieve additional evidence from sources such as PubMed. The vision agents perform tasks such as chest X-ray classification and skin-lesion segmentation, with their outputs treated as preliminary analyses that require appropriate review. The RAG agent retrieves relevant medical documents from Qdrant while preserving the associated source metadata. Local evidence can also be supplemented with PubMed results when more recent or additional information is required. Finally, the output guardrails check the generated response for unsupported certainty, unsafe treatment advice, and recommendations that could lead to delays in seeking emergency medical care.

VI. RETRIEVAL AND EVIDENCE WORKFLOW

The retrieval and evidence workflow is designed to reduce unsupported or hallucinated information. Instead of relying solely on the language model's internal knowledge, the system retrieves relevant sources, reranks the retrieved information, and links the generated response to the supporting evidence.

1. Document Ingestion

Medical documents are added to the knowledge base while retaining useful structural information such as section headings, tables, figures, page numbers, and source metadata. Docling-based parsing is particularly useful for medical PDFs because these documents often contain complex layouts, tables, and diagrams that can be lost or distorted when basic text-extraction methods are used.

2. Hybrid Retrieval and Reranking

For a user query q, the RAG module can generate a set of related queries to improve the retrieval of relevant information:

$$Q' = \{q, q1, q2, \ldots, qn\}, \ldots(2)$$

where q_i represents related medical-domain reformulations. The system can then use hybrid retrieval:

$$Score(d, q) = \alpha Sdense(d,q) + (1 - \alpha) Sspaarse\ (d, q), \quad \ldots(3)$$

Here, dense retrieval captures semantic similarity, while sparse retrieval helps identify exact terms such as drug names, abbreviations, disease labels, and laboratory terminology. A reranking stage then orders the retrieved chunks based on their relevance before passing the most useful evidence to the response generator.

3. RAG-to-Web Fallback

When the retrieved evidence is insufficient or has low confidence, the system can transfer the query to the web-search processor for additional information. This is useful because a local medical knowledge base cannot cover every possible question. In such cases, retrieving additional evidence or deferring the response is safer than generating a confident answer based on limited information.

VII. HUMAN-IN-THE-LOOP VALIDATION

Human review is an important safety mechanism for handling high-risk outputs. Medical-image results can be sent for human validation, while lower-risk tasks may be handled automatically. Uncertain or clinically sensitive results are referred to a qualified reviewer for further assessment. The interface supports this process by displaying agent results, source links, relevant disclaimers, and human-review controls. Model predictions are presented as preliminary findings rather than definitive diagnoses.

VIII. METHODOLOGY

1. Study Design and System Boundary

This study combines a repository-based system analysis with a small functional evaluation of the two implemented medical-vision agents. The analysis follows the complete workflow, from input screening and routing to specialist processing, guardrails, validation, and final response generation. The experiments were conducted directly on the specialist models rather than through the browser interface or the LLM-based router. This approach helped isolate the behavior of the vision models from factors related to routing, network communication, and the user interface.

The chest X-ray agent uses a DenseNet-121 model with the supplied covid_chest_xray_model.pth checkpoint. The input images are converted to RGB, resized to 150 × 150 pixels, and normalized using ImageNet statistics before being classified into two categories: covid19 or normal. The skin-lesion agent uses a U-Net model with the checkpointN25_.pth.tar checkpoint. Its input images are converted to RGB, scaled to the [0, 1] range, and resized to 256 × 256 pixels. The model then generates a binary lesion mask using sigmoid activation followed by a 0.5 threshold.

2. Inference Procedure

Inference was carried out on 31 August 2026 using an AMD Ryzen 5 5600H CPU with Python 3.14.7, PyTorch 2.13.0, Torchvision 0.28.0, and OpenCV 5.0.0. Both models were run in evaluation mode without gradient computation or test-time augmentation. Each input was processed once, without any manual correction or modification.

For the chest X-ray images, the model predictions were compared with the labels provided in the repository. For the skin images, we recorded whether the model completed the inference successfully along with the predicted foreground percentage. This percentage is included only as a descriptive result and should not be interpreted as a measure of segmentation accuracy.

3. Evaluation Measures

With COVID-19 treated as the positive class, accuracy, sensitivity, and specificity were calculated using the following formulas:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad \ldots(4)$$

$$\text{Sensitivity} = TP / (TP + FN), \text{Specificity} = TN / (TN + FP) \quad \ldots(5)$$

Precision and F1 score were also calculated where applicable. Since the chest X-ray evaluation included only five samples, a two-sided 95% exact binomial confidence interval was calculated for the overall accuracy.

For image segmentation, Dice and IoU are commonly calculated using the following formulas:

$$Dice = \frac{(2|P \cap G|)}{(|P| + |G|)'} \qquad IoU = \frac{(|P \cap G|)}{(|P \cup G|)} \qquad \ldots(6)$$

However, the selected skin-lesion images did not have corresponding expert-annotated masks. As a result, Dice and IoU could not be calculated. Instead, the evaluation reports whether inference was completed successfully and the percentage of the predicted mask identified as foreground.

4. Validity and Safety Controls

The model outputs were treated as technical predictions rather than clinical diagnoses. The evaluation also took into account several limitations, including the small sample size, possible overlap with development data, limited demographic information, and the absence of expert-annotated masks for the skin images. These limitations were considered when interpreting the results so that successful model execution was not presented as evidence of clinical performance.

IX. EVALUATION STRATEGY

Since the project integrates text retrieval, LLM-based reasoning, medical-image analysis, speech processing, and human validation, its evaluation should cover multiple aspects rather than depend on a single metric. Table I summarizes the recommended evaluation metrics, with an emphasis on overall system behavior rather than conversational fluency alone.

The evaluation should distinguish between retrieval quality and response quality. Reviewers can assess whether the top-k retrieved chunks provide adequate support for the generated answer and whether the cited sources support the individual claims. Routing tests should cover a range of cases, including normal conversations, health-related questions, requests for current evidence, image uploads, irrelevant images, ambiguous questions, and emergency situations. For each test case, the expected agent selection and corresponding safety action should be recorded.

Vision models should be evaluated using appropriate classification or segmentation metrics, along with tests that examine their performance under dataset shift. Safety evaluation should include measures such as escalation recall, appropriate refusal of unsafe requests, citation faithfulness, and the rate of unsafe advice. Clinical reviewers can further assess the factual accuracy of responses, important omissions, potential risks, and possible sources of bias [5].

Table I: Evaluation Criteria for the Multi-Agent Medical Assistant

Metric	Purpose
Retrieval precision	Measures whether retrieved chunks support the answer.
Citation faithfulness	Checks whether generated claims match cited sources.
Answer relevance	Measures usefulness of the final response.
Routing accuracy	Evaluates whether the correct agent receives the task.
Fallback accuracy	Measures whether weak RAG evidence triggers web search or deferral.
Vision accuracy	Measures image-classification or segmentation quality.
Escalation recall	Measures detection of high-risk cases.
Guardrail success	Tracks unsafe, biased, or unsupported responses.
User satisfaction	Captures clarity and usability.
Human-review latency	Measures delay introduced by validation.

.

X. RISK MODEL

The system can assign a risk level to each interaction based on factors such as the type of input, user intent, quality of available evidence, and potential consequences. Let Rdenote the resulting risk score:

$$\mathbf{R} = \mathbf{wmM} + \mathbf{weE} + \mathbf{wcC} + \mathbf{wuU}, \quad \ldots(7)$$

Here, M represents modality-related risk, E represents weak or insufficient evidence, C represents potential clinical consequences, and Urepresents uncertainty in the user's request. The weights should be selected conservatively. Factors such as medical-image uploads, possible emergency symptoms, low retrieval confidence, and treatment-related requests should increase $R$and make escalation more likely.

The purpose of this model is to provide a structured framework for safety rather than to define a fixed or exact risk formula. Different types of interactions may require different levels of caution. Incorporating risk into the routing process makes safety-related decisions more explicit, consistent, and easier to evaluate.

XI. RESULTS

1. Chest X-Ray Classification

Table II summarizes the five chest X-ray samples and the corresponding model predictions. All three COVID-19 images and both normal images were classified correctly. With COVID-19 considered the positive class, the results were TP = 3, TN = 2, FP = 0, and FN = 0. Based on this small pilot subset, the model achieved 100% accuracy, sensitivity, specificity, precision, and F1 score.

Although the observed accuracy was 5/5 (100%), the two-sided 95% exact binomial confidence interval was approximately 47.8%–100%. The wide interval highlights the uncertainty associated with evaluating the model on only five samples and shows that this sample size is insufficient to establish generalizable diagnostic accuracy. Moreover, the images were repository demonstration files rather than an independent test set. Therefore, these results indicate that the model and preprocessing pipeline produced the expected outputs on the selected examples, but they should not be interpreted as evidence of clinical validity.

Table II: Five-Sample Output of the Repository's Chest X-Ray Agent

Input image	Reference label	Agent output	Correct
covid-example-01.jpg	COVID-19	covid19	Yes
covid-example-02.jpeg	COVID-19	covid19	Yes

covid-19-pneumonia-101.png	COVID-19	covid19	Yes
NORMAL2-IM-0362-0001.jpeg	Normal	normal	Yes
NORMAL2-IM-0374-0001.jpeg	Normal	normal	Yes

2. Skin-Lesion Segmentation

The skin-lesion agent successfully generated a nonempty binary mask for all five selected inputs, resulting in a technical completion rate of 5/5 (100%). Table III presents the foreground area obtained from each predicted mask, which ranged from 3.2% to 62.8%. The variation across the samples shows that the model generated different mask outputs for the selected cases rather than producing a constant or identical result.

Table III: Five-Sample Output of the Repository's Skin-Lesion Segmentation Agent

Input image	Agent output	Foreground area	Inference completed
ISIC_0012558.jpg	Binary lesion mask	9.1%	Yes
ISIC_0012559.jpg	Binary lesion mask	6.7%	Yes
ISIC_0012560.jpg	Binary lesion mask	3.2%	Yes
ISIC_0012563.jpg	Binary lesion mask	62.8%	Yes
ISIC_0012564.jpg	Binary lesion mask	16.8%	Yes

Foreground area should not be interpreted as a measure of segmentation accuracy. Since the five selected images were available in the repository without corresponding expert-annotated masks, metrics such as Dice, IoU, pixel accuracy, sensitivity, and specificity could not be calculated reliably. The results only demonstrate that the U-Net checkpoint loaded successfully and produced nonempty masks for all five inputs. A proper evaluation of segmentation performance would require a held-out dataset with expert-annotated masks, allowing metrics such as Dice and IoU to be calculated using the predefined threshold.

Overall Evaluation

All ten model inferences across the two tasks were completed successfully without any failures. The chest X-ray model matched all five repository-provided labels, while the skin-lesion model generated five valid, nonempty masks without providing a quantitative accuracy estimate. These findings demonstrate that the models and inference pipelines operated successfully on the selected examples, but they do not establish diagnostic or segmentation accuracy. Therefore, the system is better described as an operational research prototype rather than a clinically validated system, and these results should not be used for direct comparison with validated medical-imaging systems.

XII. CONCLUSION

This paper presented the Multi-Agent Medical Assistant as a workflow-based medical AI assistant that integrates evidence retrieval, agent routing, medical-image processing, guardrails, and human validation. In the ten-inference pilot, the chest X-ray agent matched all five repository-provided labels, while the skin-lesion agent generated nonempty masks for all five selected inputs. However, the small sample size, possible dependence on repository data, and absence of expert reference masks limit the conclusions that can be drawn about external accuracy. The findings highlight that responsible medical AI involves more than generating fluent responses; it also requires traceable evidence, appropriately bounded claims, clear escalation mechanisms, human review, and independent evaluation.

Future work should focus on evaluation using validated medical datasets, improved multilingual support, formal safety testing, external medical-imaging benchmarks, and greater involvement of clinical experts. Additional work could also explore privacy-preserving logging and evaluate the system in collaboration with healthcare professionals to better understand its practical use and limitations.

Acknowledgment

The authors sincerely thank Dr. K. Himabindu, Associate Professor, Department of Computer Science & Engineering,

PIET, Parul University, for her valuable guidance, continuous support, and insightful suggestions throughout this research.

REFERENCES

1. Papiya Mahajan, Rinku Wankhade, Anup Jawade, Pragati Dange, and Aishwarya Bhoge, "Healthcare chatbot using natural language processing," International Research Journal of Engineering and Technology, vol. 7, no. 11, pp. 1715–1720, 2020.
2. M. V. Patil, Subhawna, Priya Shree, and Puneet Singh, "AI based healthcare chat bot system," International Journal of Scientific and Engineering Research, vol. 12, no. 7, pp. 668–671, 2021.
3. Harsh Mendapara, Suhas Digole, Manthan Thakur, and Anas Dange, "AI based healthcare chatbot system by using natural language processing," International Journal of Scientific Research and Engineering Development, vol. 4, no. 2, pp. 89–96, 2021.
4. Jagbeer Singh, Vaibhav Deshwal, Sourabh Kumar, Manish Khaloria, Manish Yadav, and Priyanshu Negi, "A healthcare chatbot system using python and NLP," Journal of Pharmaceutical Negative Results, vol. 13, no. Special Issue 10, pp. 5528–5536, 2022.
5. Karan Singhal, Shekoofeh Azizi, Tao Tu, et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172–180, 2023.
6. Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz, "Capabilities of GPT-4 on medical challenge problems," arXiv preprint arXiv:2303.13375, 2023.
7. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
8. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2024.
9. N. Athulya, K. Jeeshna, S. J. Aadithyan, U. Sreelakshmi, and Hairunizha Alias Nisha Rose, "Healthcare chatbot," International Journal of Creative Research Thoughts, vol. 9, no. 10, pp. 65–70, 2021.
10. Niraj A. Wanjari, Roshan S. Ghosare, Shubham K. Dhote, Ashwini P. Meshram, and Rahul Bhandekar, "AI healthcare chatbot using natural language processing," International Research Journal of Modernization in Engineering Technology and Science, vol. 4, no. 5, pp. 1219–1222, 2022.
11. G. V. V. Sai Roshan, V. Koushik Raj, D. Mohan Satyendra, et al., "Healthcare chatbot using natural language processing," Tuijin Jishu / Journal of Propulsion Technology, vol. 45, no. 2, pp. 1829–1835, 2024.
12. Deepshika Pillai and S. R. Nalamwar, "Healthcare chatbot using natural language processing," Alochana Journal, vol. 13, no. 5, pp. 211–221, 2024.
13. Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao, "ReAct: Synergizing reasoning and acting in language models," in International Conference on Learning Representations (ICLR), 2023.
14. S. Saeedi, S. Rezayi, H. Keshavarz, et al., "MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques," BMC Medical Informatics and Decision Making, vol. 23, p. 16, 2023.
15. B. Babu Vimala, S. Srinivasan, S. K. Mathivanan, et al., "Detection and classification of brain tumor using hybrid deep learning models," Scientific Reports, vol. 13, p. 23029, 2023.
16. M. Z. Khaliki and M. S. Başarslan, "Brain tumor detection from images and comparison with transfer learning methods and 3-layer CNN," Scientific Reports, vol. 14, p. 2664, 2024.
17. Joanne Cleverley, James Piper, and Melvyn M. Jones, "The role of chest radiography in confirming COVID-19 pneumonia," BMJ, vol. 370, p. m2426, 2020.
18. R. Yasin and W. Gouda, "Chest X-ray findings monitoring COVID-19 disease course and severity," Egyptian Journal of Radiology and Nuclear Medicine, vol. 51, p. 193, 2020.
19. D. Cozzi, M. Albanesi, E. Cavigli, et al., "Chest X-ray in new coronavirus disease 2019 (COVID-19) infection: Findings and correlation with clinical outcome," La Radiologia Medica, vol. 125, pp. 730–737, 2020.
20. R. Jain, M. Gupta, S. Taneja, et al., "Deep learning based detection and analysis of COVID-19 on chest X-ray images," Applied Intelligence, vol. 51, pp. 1690–1700, 2021.

21. E. M. F. El Houby, "COVID-19 detection from chest X-ray images using transfer learning," Scientific Reports, vol. 14, p. 11639, 2024.
22. Noel C. F. Codella, David Gutman, M. Emre Celebi, et al., "Skin lesion analysis toward melanoma detection," arXiv preprint arXiv:1710.05006, 2017.
23. Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, et al., "A survey on deep learning for skin lesion segmentation," Medical Image Analysis, vol. 88, p. 102863, 2023