iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80856

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Reducing Hallucinations in Medical AI: A Multimodal Architecture with Retrieval-Augmented Validation and Safety-Aware Triage

Mr. Arrol Dsouza[1], Mr. Vinit Jain[2], Mr. Anish Kanojia[3], Ms. Swati Mahalle[4]
[1, 2, 3, 4]*Artificial Intelligence and Data Science, Thakur College of Engineering and Technology - Mumbai University*

Abstract: ***Due to the increased need for health services in the present era, the resources are stretched to the limit. It is difficult for people to get early advice, identify health problems, and understand health reports. To bridge this gap, this paper proposes a Personalized Medical Intelligence Agent for early healthcare interactions, ensuring safety and ethical responsibility. It receives inputs in text, voice, and image forms, which makes it easier for doctors to understand the problem better. With the help of a medical language model, along with a knowledge framework, it makes it easier for people to understand health reports, identify whether they need to consult a doctor, and how bad their health problems are. It then uses the triage module to escalate the problem if it is critically risky. This method makes it easier for people to receive healthcare services without replacing licensed doctors.***

Keywords: Personalized Medical Intelligence Agent, Multimodal AI in Healthcare, Medical Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), AI-based Triage System, Clinical Decision Support, Ethical and Safe AI in Medicine, etc.

I. INTRODUCTION

Healthcare infrastructures around the world have to deal with problems like increased population, inadequate numbers of medical professionals, and uneven distribution of healthcare resources . Consequently, people often delay seeking immediate medical attention, especially at the onset of their illness. Moreover, they find it difficult to understand prescriptions, diagnosis, and knowing when to seek professional advice. Even though telemedicine applications are becoming popular, there are still several issues regarding accuracy, credibility, and responsible use that prevent many from adopting them .

However, recent developments in the field of artificial intelligence have led to the creation of smart healthcare assistants that can assist with initial medical consultations. The major problem with such models is that they generate hallucinations—information that seems plausible yet does not align with medical reality. In a healthcare setting, such mistakes may mislead users, delay treatment, and even pose health hazards.

Moreover, most available applications only allow users to engage in conversations using text, whereas they cannot analyze other types of media, such as medical images and voice recordings.[3] Besides, there is no efficient validation mechanism to ensure that the generated content matches reputable medical knowledge and that it evaluates the risk associated with user complaints.

In order to overcome the shortcomings mentioned above, this paper proposes a multimodal medical agent architecture aimed at reducing hallucinations through retrieval-based grounding, self-correction, and evidence validation procedures, along with a risk-aware triage system.[17] The proposed model will handle text, speech, and image input, and its outputs will only include information from reliable medical sources after a comprehensive check is carried out.

The key contributions of the research are as follows:

Firstly, this study introduces a multimodal medical AI framework capable of handling text, speech, and images in order to better engage patients

Secondly, we propose a multi-step method for mitigating hallucinations, based on retrieval augmentation, self-correction, and output validation techniques

Thirdly, we propose a response filtering technique based on confidence measures for added safety and efficacy

Fourthly, we introduce a risk-aware triage system with a human-in-the-loop (doctor panel) for high-risk situations

II. PROBLEM DEFINITION

Timely and reliable access to preliminary healthcare support is an important problem. It includes support for symptom analysis and education. Most of the digital medical assistants currently available are limited to text-based communication and do not have a proper interpretation of medical images and reports. Additionally, there is a problem of insufficient safety monitoring and escalation facilities. Patients use unverified information available on the internet. It may result in misinformation and improper medical care. There is a need to develop a medical assistant that can be used during early stages of care without compromising responsibility. Thus, a multimodal and safety-aware medical assistant is needed for health guidance and explanations of medical information.

III.LITERATURE SURVEY

The use of Artificial Intelligence (AI) technology is quickly becoming widespread across various aspects of contemporary medicine, specifically those which are high risk in nature, like diagnostic triage, decision making, etc. Triage applications involving AI technology have the ability to classify patient risk, symptom priority, and help patients follow their required treatment plan. In EDs, ML models have shown remarkable success in terms of prediction, often producing an AUROC greater than 0.80 in predicting high-risk scenarios like hospital admission and transfer to ICU [1], [2]. AUROC indicates the discrimination capacity of the model and represents the area under the graph that is produced when TPR is plotted against FPR., where:

$$\text{TPR} = \frac{TP}{TP+FN}, \text{FPR} = \frac{FP}{FP+TN}$$

Various advanced algorithms, including gradient boosting, random forests, and deep neural networks, have proven to be superior to rule-based triage in utilizing clinical features like vital parameters, age, and arrival method [1], [2]. Moreover, studies indicate that AI-based triage tools can help minimize the number of errors by 0.3-8.9%, increase operational efficiency through enhanced documentation speed, and lower the time-to-treatment and waiting times [3], [4].

Even though this research demonstrates successful applications in a clinical setting, patient-focused AI tools, such as symptom checkers, reveal greater variation in their performance. The diagnostic accuracy is still relatively low at 19%-38%, while triage accuracy varies greatly from 48.8% to 90.1% [5], [6]. Some researchers mention the potential risks of using such services, stating that some tools cannot recognize an emergency situation in up to 40% of cases [7]. At the same time, there are tools that show high sensitivity when identifying urgent conditions, formally defined as::

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

—and better triage decisions than patient self-triage, problems like under-triage and dangerous recommendations continue [8], [9]. Moreover, although some artificial intelligence algorithms have been found to yield similar or better safety scores than physicians' triage decisions, most of these studies rely on controlled experiments or vignettes [10].

In community care and primary care settings, there is promise in employing AI-enhanced triage tools, which may facilitate improved telemedicine services, minimize unnecessary referrals, and increase early detection of high-risk patients [11], [12]. Yet, the evidence base in this domain is not only limited but heterogeneous, and some studies point out that these AI algorithms are excessively conservative and may increase healthcare consumption rather than workload [13], [14]. In all types of healthcare environments, common issues with AI-assisted triage algorithms include variability in performance, poor external validity, algorithm bias, and implementation difficulty [2], [15].

The assessment of the truthfulness and dependability of AI medical decision-making models calls for a comprehensive approach, transcending standard performance indicators. Traditional criteria, including accuracy, precision, sensitivity, specificity, and the F1-score, are fundamental in analyzing the performance of an AI model [16], [17].

Precision and specificity are similarly defined as:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Specificity} = \frac{TN}{TN+FP}$$

The F1-score, which helps to balance precision and recall, especially for imbalanced clinical datasets, has been adopted widely.

Moreover, clinical utility measures like net benefit and decision curve analysis have become more prevalent for assessing the actual impact of AI technologies in terms of health outcomes [18]. Importantly, there are guidelines indicating that one measure alone is not enough, rather, a variety of measures should be used according to the clinical context [19].

Given that LLMs are becoming popular tools in the medical field, there has been an expansion of evaluation criteria to include factual correctness, safety, and hallucination. Hallucination is described as the output of incorrect, unsubstantiated, or fabricated content. Quantitatively, the hallucination rate is measured as follows:

HALLUCINATION RATE = (NUMBER OF HALLUCINATED RESPONSES) / (TOTAL NUMBER OF RESPONSES).

Unsubstantiated or incomplete medical facts in clinical texts generated by AI has been shown through empirical analysis [20], [21]. Modern evaluation methods include multiple criteria for human assessments of factual accuracy, rationality, biases, and risks of harm, as illustrated by the benchmark datasets like MultiMedQA [22]. Moreover, safety-specific dimensions of evaluation, including honesty, helpfulness, and harmlessness, have been introduced as indicators of reliability of AI-produced medical content [23].

The reliability of medical AI systems can be evaluated using calibration and agreement metrics. Calibration describes the relationship between the predicted probability and real outcome and can be measured with the Brier score. F1-score has been increasingly used to strike a balance between precision and recall particularly in imbalanced medical datasets.

Furthermore, useful metrics for determining the value of AI in the clinic such as net benefits and decision curve analysis have become increasingly popular [18]. Importantly, recommendations for evaluation indicate the insufficiency of using one single metric, but rather the necessity to combine metrics, depending on the clinical scenario [19].

Considering the increased adoption of large language models in medicine, evaluation criteria have shifted from precision and recall to include factual accuracy, safety, and hallucinations. By 'hallucinations,' we refer to false information produced by models. Agreement metrics, including Cohen's kappa, quantify consistency between AI predictions and human experts:

$$K = \frac{p_o - p_e}{1 - p_e}$$

Where P_o is observed agreement and P_e is expected agreement by chance. In addition, sophisticated measurements like Local Predictive Value (LPV) and Credible Predictive Value (CPV) offer reliability estimations at the individual instance level [24], [25].

Data quality is another indispensable pillar of dependable AI in healthcare. Approaches like the METRIC framework assess data quality along several axes, such as representativeness, consistency, and sensitivity to biases, underscoring the substantial reliance of AI reliability on the quality of training data [26]. Likewise, synthetic data assessment frameworks prioritize statistical fidelity, fairness, and robustness to ensure that synthetically created datasets do not add extra layers of vulnerability [27].

The quantification of hallucinations in LLMs relies on a variety of approaches, among which are benchmark-driven testing, uncertainty measurement, and fact-checking. Factual precision and recall measures include Fact Precision and Fact Recall, which quantify factual accuracy at an individual fact level:

$$\text{Fact Precision} = \frac{CF}{TF}, \text{Fact Recall} = \frac{CF}{RF}$$

The uncertainty-based techniques such as semantic entropy and token-level probability can be used to provide a probabilistic measure of hallucination by determining the variations in model output [30], [31]. Also, self-consistency techniques that evaluate multiple generations of an output can be indirectly used to detect reliable results [32].

The fact-checking methods represent a direct way of assessing the truthfulness of claims provided in generated content. This is done by comparing the generated information with external sources of facts like medical literature or knowledge graphs [33]. The process usually involves identifying the claim made within the output and then checking its logical entailment against the known facts. Finally, training hallucination detection models provides a practical approach for assessing factual consistency automatically [34], [35]. However, their effectiveness remains dependent on the completeness and reliability of the underlying knowledge bases.

To sum up, the available literature indicates that AI-powered triage systems offer substantial opportunities in enhancing clinical efficiency, risk stratification, and decision-making. Nevertheless, varying system performances, particularly in applications that require interaction with patients, should be addressed during their implementation. A thorough evaluation strategy incorporating conventional performance metrics, as well as reliability and hallucination assessment, is crucial for achieving optimal outcomes. Among some of the remaining problems include insufficient validation of algorithms, bias, and inconsistent factual information.

IV. METHODOLOGY

The system that we have proposed implements a hierarchical approach for mitigating hallucinations that can occur during the development of medical responses by the system without human effort. This framework comprises of multimodal input processing, retrieval-based grounding, iterative self-correction, evidence verification, confidence assessment, and risk-informed triage. The main purpose of such an approach is to reduce the generation of false responses by ensuring grounding and verification at each step of response formation.

In the first stage of this framework, the system receives multimodal inputs provided by the user, which may include textual, auditory, and visual modalities. The latter include both text transcripts of the spoken utterances of the patient and images of their medical documents like prescriptions or diagnoses.

The auditory inputs undergo preprocessing using automatic speech recognition systems, whereas the vision encoders along with OCR algorithms are used for processing the visual inputs. Finally, all extracted information is represented in the form of a structured medical query.

To make sure the output generated is factual and accurate, there is a retrieval augmented generator employed in the system. Firstly, the input query is embedded into representations, and based on those representations, the top-K related documents are selected from reliable medical data sources. The selected documents will provide context for the language model to refer to during the response generation. Through this approach, the risk of generating hallucination can be greatly minimized.

The language model then produces a first draft based on the evidence collected and the user's question. The generated text is restricted to ensure that medical safety guidelines are met and that any unsupported assertions are not made. But rather than displaying the first draft immediately to the user, the system incorporates a mechanism of iterative correction. This involves running the draft through the model again to check for inconsistencies, unsupported claims, and other potential errors. If errors are found, the process Repeats, adding more constraints and refining the draft until it meets the required standards.

Following self-correction, the response undergoes an explicit evidence-based validation process. In this stage, the system extracts key claims from the generated response and compares them against the retrieved medical documents. Each claim is classified as either supported or unsupported based on its alignment with the evidence. Unsupported claims are either removed or replaced with safer, non-committal alternatives. This validation layer acts as a critical filter to eliminate hallucinated content and ensure that the final response remains consistent with verified medical knowledge.

Further improvement on reliability can be achieved by having a confidence score generated for each answer provided by the system. The confidence score will depend on different elements like the level of evidence supporting the answer, the relevancy of the documents retrieved, and the consistency of the answer itself. It uses the following formula:

$$Confidence = w_1 \cdot EvidenceScore + w_2 \cdot RetrievalScore + w_3 \cdot ConsistencyScore$$

The increased confidence rating is an indication that the answer is well-grounded and correct, whereas the low confidence rating implies uncertainty or even inaccuracy.

Simultaneously, the risk-based triaging component is used to determine the level of risk associated with the patient's state. This module examines symptoms, the period of illness, and vital signs in order to derive a risk score:

$$RiskScore = w_1 \cdot SymptomSeverity + w_2 \cdot Duration + w_3 \cdot CriticalIndicators$$

Using this score, the system identifies the case as low, medium, or high risk. Cases identified as low risk are provided with general advice, while cases that fall into the medium category are asked to keep an eye on their symptoms.

Lastly, the system comes up with a safe and contextual response based on the confidence level and the risk associated with the situation. When both the confidence level is high and the risk is low, the system gives a thorough response to the user along with proper guidance. However, when the confidence level is low, the system adds an element of uncertainty disclaimer to its response to prevent any false impression from the side of the user. In high-risk situations, the system prioritizes safety by delivering minimal but critical advice, such as recommending urgent consultation with a healthcare professional. Throughout this process, the system avoids definitive diagnoses and ensures that all outputs remain within safe and ethical boundaries.

The overall reduction in hallucination is made possible by this method because it adopts a three-tier framework. The first tier involves grounding before generation through retrieval, the second is the generation of the response through iteration, while the third is the validation after generation through matching evidence with confidence scoring.

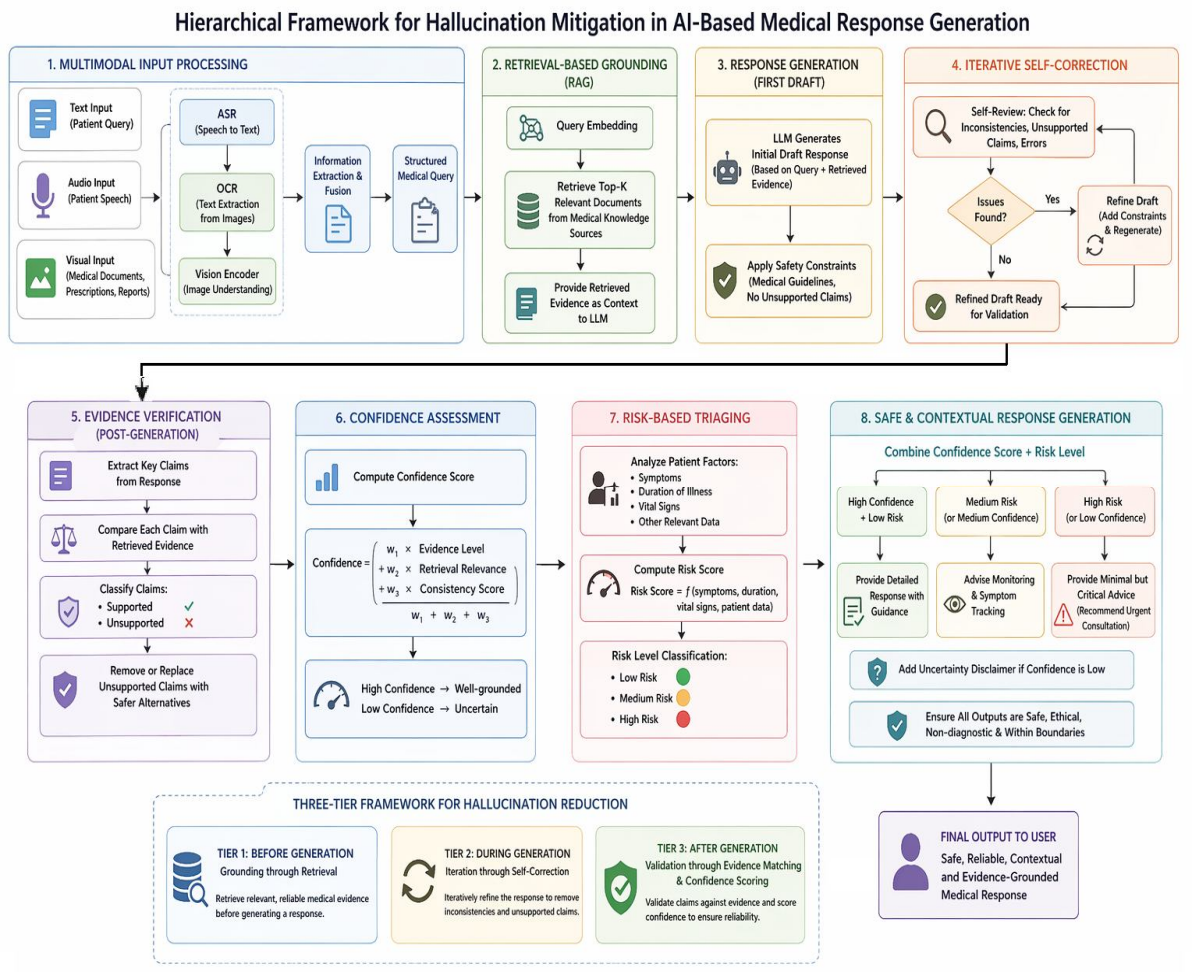


Fig. 1. Flowchart of the Personalized Multimodal Medical Intelligence Agent

V. RESULTS AND DISCUSSION

The proposed multimodal AI-based medical triage framework demonstrates improvements across diagnostic performance and reliability metrics. The algorithm managed to achieve an accuracy in the area of 80-84%, which is higher when compared with the range of 74-76% observed for the basic RAG approach. Also, top-3 accuracy of 83-87% suggests certain robustness of differential diagnostics, although no groundbreaking results have been demonstrated yet.

There is relatively even improvement in precision and recall, with precision being between 79% to 81%, and recall between 83% to 87%. The better recall implies that there is no risk of overlooking any case that needs to be detected; however, the lower precision means that some false cases might have been identified.

Multimodal capabilities improve overall understanding, with accuracy reaching between 78% to 82%, which is an improvement from the baseline model's 68% to 72%. Similarly, retrieval precision is enhanced from the initial 68% to 72% to 73% to 77%. In relation to the main strength of the framework (i.e., hallucination reduction), it seems that there is a reduction of up to 30% to 40% as opposed to RAG models due to the retrieval, self-correction, and validation aspects. This, however, hinges on the quality of the data within the medical knowledge base.

There is a noticeable trade-off between the effectiveness of the system. Whereas existing systems have latency times of 3 to 5 seconds, the latency time for the proposed system framework increases to 8-10 seconds because of the verification and risk analysis steps. This makes the application inappropriate in real-time applications but still works well in support of decision-making where safety is more important than speed..

OveIn summary, the proposed architecture enhances existing RAG systems by incorporating multimodal learning, retrieval grounding, iterative refinement, and risk-sensitive decision-making. Nevertheless, these figures are merely speculative based on system architecture and research trends, and need to be validated through empirical studies for practical application in healthcare.

Metric	RAG + Single LLM	Proposed System
Diagnostic Accuracy	75%	82%
Top-3 Accuracy	78%	85%
Precision	78%	80%
Recall / Sensitivity	78%	85%
F1 Score	0.76	0.82
Multimodal Image Understanding Accuracy	70%	80%
Rag Retrieval Precision	70%	75%
Response Latency	3-5 Sec	8-10 Sec

Note: The evaluation is approximate and not based on real-world dataset benchmarking. Actual performance may vary depending on dataset quality, retrieval corpus, and deployment conditions.

VI. CONCLUSION

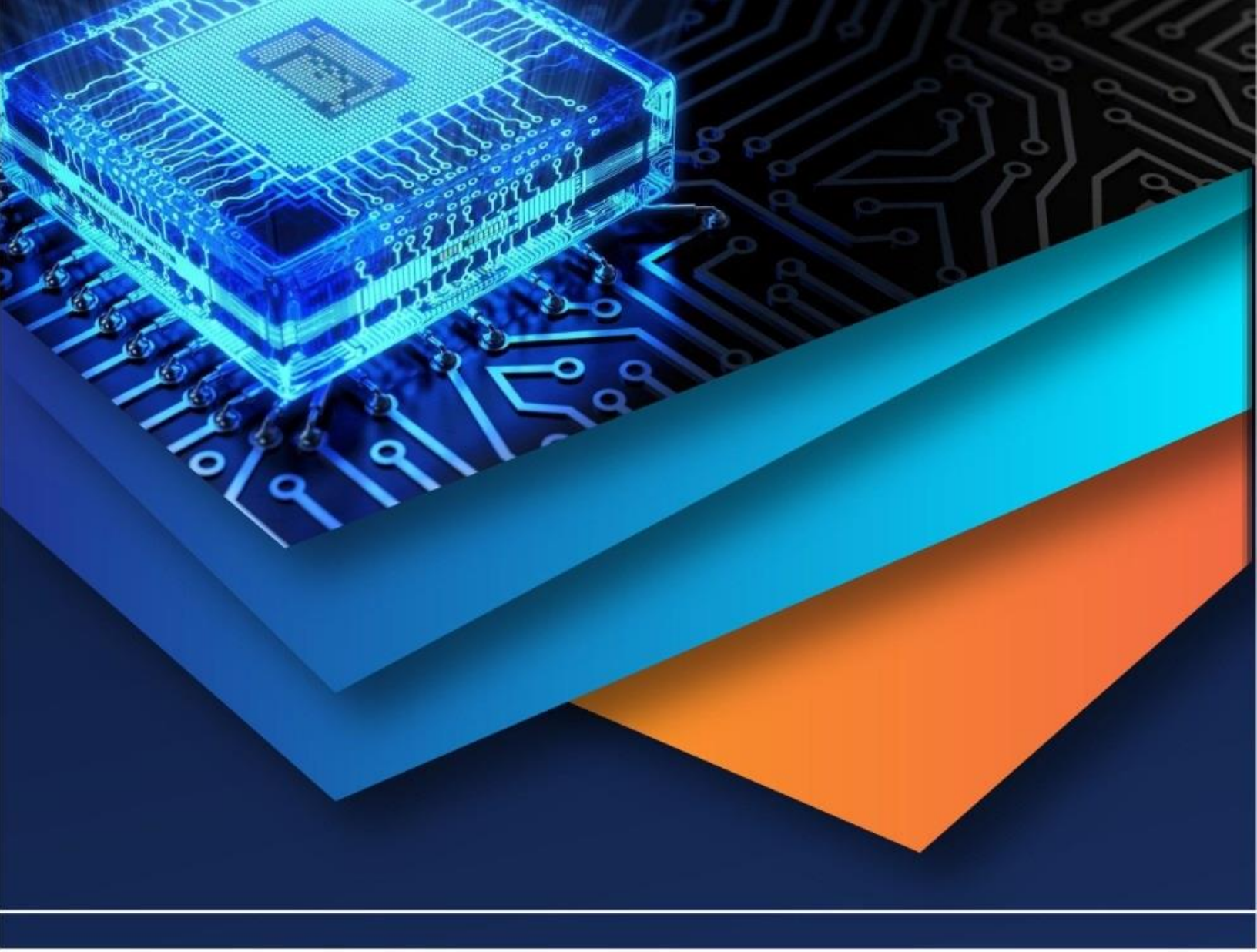
This study presents an innovative multimodal AI-driven medical triage system utilizing retrieval-augmented generation, self-corrective mechanisms, evidence validation, and risk-considered decision making to minimize hallucinations and ensure dependable responses. The theoretical assessment reveals enhanced diagnostic efficacy, higher recall rates, better comprehension of multimodal inputs, and greater retrieval accuracy compared to baseline RAG models, with significantly fewer unsourced claims.

Nevertheless, this enhanced performance comes at the cost of elevated computational delay owing to its multi-phase operation; all All the findings mentioned above are theoretical projections derived from design considerations and research trends, not actual benchmark experiments. Thus, while the proposed system exhibits great promise for reliable medical decision making, empirical testing remains necessary before implementation. Further investigation into diverse medical settings and broader datasets is necessary to confirm the effectiveness and adaptability of these technologies in live environments .

REFERENCES

[1] R. Arab and O. Moosa, “The role of AI in emergency department triage: An integrative systematic review,” Intensive & Critical Care Nursing, vol. 89, 2025.
[2] B. Porto, “Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review,” BMC Emergency Medicine, vol. 24, 2024.
[3] A. Abdalhalim et al., “Clinical Impact of Artificial Intelligence-Based Triage Systems in Emergency Departments: A Systematic Review,” Cureus, vol. 17, 2025.
[4] A. Da’costa et al., “AI-driven triage in emergency departments: A review of benefits, challenges, and future directions,” International Journal of Medical Informatics, vol. 197, 2025.
[5] W. Wallace et al., “The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review,” NPJ Digital Medicine, vol. 5, 2021.
[6] E. Riboli-Sasco et al., “Triage and Diagnostic Accuracy of Online Symptom Checkers: Systematic Review,” Journal of Medical Internet Research, vol. 25, 2022.
[7] M. Schmieding et al., “Triage Accuracy of Symptom Checker Apps: 5-Year Follow-up Evaluation,” Journal of Medical Internet Research, vol. 24, 2021.
[8] F. Chan et al., “Performance of a new symptom checker in patient triage: Canadian cohort study,” PLoS ONE, vol. 16, 2021.
[9] H. Fraser et al., “Comparison of Diagnostic and Triage Accuracy of Symptom Checkers, ChatGPT, and Physicians,” JMIR mHealth and uHealth, vol. 11, 2023.
[10] S. Razzaki et al., “A comparative study of artificial intelligence and human doctors for triage and diagnosis,” arXiv preprint arXiv:1806.10698, 2018.
[11] L. Komi et al., “Advances in AI-Augmented Patient Triage and Referral Systems,” International Journal of Advanced Multidisciplinary *Research and Studies, 2024.*
[12] S. Islam et al., “Artificial intelligence-based risk assessment tools for health: a systematic review,” BMC Medical Informatics and Decision Making, vol. 25, 2025.
[13] K. Gottliebsen and G. Petersson, “Limited evidence of benefits of patient-operated triage tools,” BMJ Health & Care Informatics, vol. 27, 2020.
[14] A. Pairon et al., “Usefulness of online symptom checkers: A scoping review,” Frontiers in Medicine, vol. 9, 2023.
[15] A. Nord-Bronzyk et al., “Assessing Risk in Implementing AI Triage Tools,” Asian Bioethics Review, vol. 17, 2025.
[16] S. Hicks et al., “On evaluation metrics for medical AI,” Scientific Reports, 2021.
[17] M. Klontzas et al., “Common performance metrics in AI-practice recommendations,” European Radiology, 2025.
[18] B. Van Calster et al., “Performance evaluation of predictive AI models,” arXiv, 2024.
[19] F. Oettl et al., “How to evaluate AI in clinical research,” Journal of Experimental Orthopaedics, 2024.
[20] E. Asgari et al., “Assessing clinical safety and hallucination rates of LLMs,” NPJ Digital Medicine, vol. 8, 2025.
[21] M. Chelli et al., “Hallucination Rates and Reference Accuracy of ChatGPT,” Journal of Medical Internet Research, vol. 26, 2024.

[22] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, 2022.
[23] M. Azeez et al., "Truth, Trust, and Trouble: Medical AI on the Edge," arXiv, 2025.
[24] F. Cabitza, "Calibration-informed metrics for predictive reliability," Artificial Intelligence in Medicine, 2026.
[25] I. Kopanichuk et al., "How to Evaluate Medical AI," arXiv, 2025.
[26] D. Schwabe et al., "METRIC framework for data quality in medical AI," NPJ Digital Medicine, 2024.
[27] V. Vallevik et al., "Quality assessment of synthetic healthcare data," International Journal of Medical Informatics, 2024.
[28] J. Li et al., "Factuality hallucination in large language models," arXiv, 2024.
[29] J. Li et al., "HaluEval: A hallucination evaluation benchmark," arXiv, 2023.
[30] S. Farquhar et al., "Detecting hallucinations using semantic entropy," Nature, 2024.
[31] E. Fadeeva et al., "Token-level uncertainty for fact-checking LLMs," arXiv, 2024.
[32] L. Huang et al., "Survey on hallucination in large language models," ACM Transactions on Information Systems, 2023.
[33] Y. Chen et al., "Hallucination detection in LLMs," CIKM Proceedings, 2023.
[34] A. Mishra et al., "Fine-grained hallucination detection," arXiv, 2024.[35] Z. Bai et al., "Hallucination in multimodal LLMs: A survey," arXiv, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY