International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multimodal Clinical Query Analysis Using Groq-LLaMA Vision Models and Reinforcement-Inspired Evaluation

Ramya B N[1]*, Kushal S*[2]*, Kiran S Nair*[3]*, Asritha Y*[4]*, Prajwal V*[5]

[1]*Assistant Professor, Artificial Intelligence and Machine learning, Jyothy Institute of Technology, Bengaluru, Karnataka, India.*
[2,3,4,5]*Student, Artificial Intelligence and Machine learning, Jyothy Institute of Technology, Bengaluru, Karnataka, India.*

ABSTRACT

This project presents an AI-powered medical diagnostic chatbot that utilizes multimodal input for enhanced healthcare assistance. By integrating Groq-hosted LLaMA Vision models with a natural language interface, the system supports both image-based diagnosis and text-based symptom evaluation. Users can upload medical images or enter descriptive queries, which are processed by two distinct LLaMA-based models (11B and 90B), offering multiple diagnostic perspectives. The backend is developed using FastAPI, while the frontend is styled with Tailwind CSS, ensuring a seamless and interactive user experience. A reinforcement-inspired evaluation framework is employed to assess the system's performance in terms of diagnostic accuracy, treatment suggestions, and reasoning quality. Preliminary results demonstrate the effectiveness of the system in interpreting visual symptoms, providing accurate diagnostic support, and delivering an intuitive interface for diverse medical input types.

Keywords: AI Medical Chatbot, Vision-Language Model, Groq, LLaMA, FastAPI, Multimodal Diagnosis, ReinforcementInspired Evaluation, Medical Image Analysis.

1. INTRODUCTION

The increasing demand for accessible and reliable healthcare solutions has fueled the exploration of artificial intelligence (AI) in medical diagnostics. Many individuals, particularly those in remote or underserved regions, encounter significant challenges in consulting healthcare professionals promptly due to geographical constraints, time limitations, or limited access to medical resources. Early diagnosis and timely guidance are essential to prevent the progression of common ailments and to alleviate patient anxiety.

Traditional diagnostic processes often rely on face-to-face consultations, which can be time-consuming and dependent on the availability of healthcare providers. However, with recent advancements in AI, particularly in the field of deep learning, intelligent systems have emerged that can simulate clinical interactions, offer preliminary diagnostic assessments, and even interpret medical imagery. Vision-language models (VLMs), which integrate visual processing with natural language comprehension, have proven to be particularly effective in supporting multimodal healthcare applications.

This study introduces an AI-based medical chatbot designed to deliver preliminary diagnostic insights using both image and text inputs. The system employs two LLaMA-based vision-language models (LLaVA) to evaluate medical images uploaded by users and respond intelligently to symptom descriptions through natural language interaction. The backend infrastructure is developed using FastAPI, while the frontend leverages Tailwind CSS to ensure a smooth and responsive user interface. By supporting multimodal input, the system aims to enhance the accessibility and efficiency of early diagnostics, empowering users to make more informed healthcare decisions.

2. LITERATURE SURVEY

Sl no	Author(s)	Title	Source	Year	Key Contribution
1	H. Liu, Y. Zhang, Y. Du, Z. Yang	LLaVA: Large Language-and-Vision Assistant	arXiv	2023	Introduced LLaVA, combining image and text for AI assistance.
2	H. Touvron et al.	LLaMA: Open and Efficient Foundation Language	arXiv	2023	Presented scalable LLaMA models for efficient NLP tasks

3	A. Radford et al.	Learning Transferable Visual Models from Natural Language Supervision	arXiv	2021	Proposed CLIP model for joint vision-language learning.
4	D. Silver, R. Sutton, M. Müller	Reinforcement learning: A survey	Found. & Trends in ML	2008	Comprehensive overview of reinforcement learning methods.
5	A. Esteva et al.	A guide to deep learning in healthcare	Nature Medicine	2019	Explored applications of deep learning in healthcare.

3. METHODOLOGY

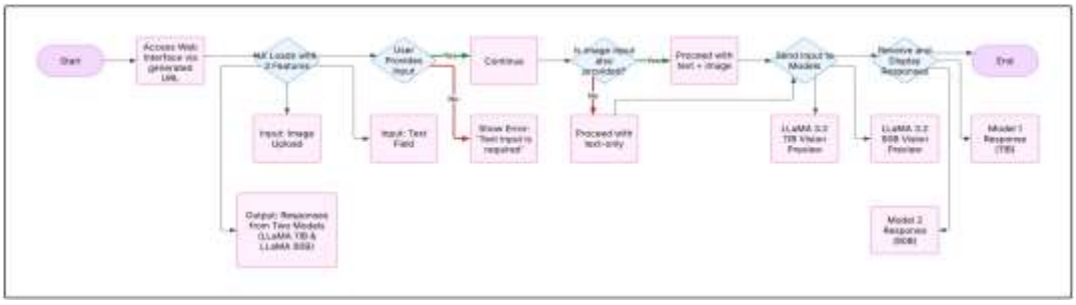
This research proposes a multimodal AI-based medical chatbot that integrates image-based diagnostics with natural language interaction to offer preliminary medical guidance. The methodology is structured around five primary components: user interaction through a web interface, input validation, vision-language model invocation, comparative response analysis, and result visualization. The system architecture is illustrated in Figure 1.

The chatbot allows users to provide inputs in two formats: (i) textual symptom descriptions or (ii) a combination of text and uploaded medical images. When a user accesses the web interface (via a generated URL), a responsive UI is loaded with three core input features: a text field, image upload option, and model toggle switch. Input validation rules enforce mandatory text submission, with optional image support. If an image is uploaded alongside the query, the system processes both through the backend pipeline.

Images are verified and preprocessed using the Python Imaging Library (PIL), ensuring consistent quality by converting formats to RGB, resizing to 224×224 pixels, and filtering out unsupported or low-quality uploads (e.g., blurry, poorly lit, or AI-generated content). Acceptable formats include JPEG, PNG, and WEBP. The backend is developed with FastAPI to manage multipart inputs, validate requests, and handle asynchronous model queries to Groq's high-performance endpoints.

Two powerful vision-language models—LLaMA 3.2 11B Vision Preview and LLaMA 3.2 90B Vision Preview—are queried in parallel with the processed input. The backend submits the text (and image, if available) to both models, and captures their respective responses. Each model's output is returned to the frontend and displayed side by side for diagnostic comparison. The system's diagnostic flow, as depicted in Figure 1, begins at user input and branches depending on whether image input is included. All valid inputs are forwarded to the models, and responses are rendered in the UI with appropriate visual distinctions between the 11B and 90B model outputs. This dual-model comparison allows users to observe variations in diagnosis and recommendations, promoting transparency in AI-driven assessments.

The frontend, styled with Tailwind CSS, ensures accessibility and visual clarity across devices. Features include drag-anddrop image upload, real-time response rendering, and Enter-key support for quick submissions. The chatbot's design prioritizes speed, user experience, and medical accuracy.


Start
Access Web Interface via generated URL
UI Loads with 3 Features
User Provides Input
Yes
Continue
No
Input: Image Upload
Input: Text Field
Show Error: "Text input is required"
Is image input also provided?
Yes
Proceed with text + image
No
Proceed with text-only
Send Input to Models
Retrieve and Display Responses
End
LLaMA 3.2 11B Vision Preview
LLaMA 3.2 90B Vision Preview
Model 1 Response (11B)
Model 2 Response (90B)
Output: Responses from Two Models (LLaMA 11B & LLaMA 90B)

4. MODELING AND ANALYSIS

4.1 Evaluation Methodology

To assess the performance of our medical chatbot models, we designed a flexible, human-interpretable scoring framework, particularly suited for open-ended natural language responses. Traditional classification metrics such as precision, recall, and F1-score were deemed inadequate due to the inherently subjective and varied nature of medical dialogue, where multiple valid diagnoses and treatments may exist for the same query.

Instead, each chatbot response was manually evaluated against three essential criteria: the presence of a diagnosis, the suggestion of a treatment, and the provision of reasoning or explanation.

Criteria Met	Score	Classification
Diagnosis + Treatment + Reason	1.0	Full Pass
Any 2 of the 3	0.66	Partial Pass
Only 1 of the 3	0.33	Minimal Pass
None / Irrelevant / Incorrect Response	0.0	Fail

This methodology balances the ambiguity inherent in free-text medical responses with the need for structured evaluation, ensuring a more nuanced analysis compared to rigid classification metrics.

Justification for Scoring Design

The chosen thresholds (1.0, 0.66, 0.33, and 0.0) effectively differentiate between fully accurate, partially correct, minimally useful, and completely unhelpful answers. More granular alternatives (such as 100–60–30–0) were considered but rejected to avoid unnecessary complexity without meaningful benefit in this setting.

4.2 Results and Observations

The evaluation was conducted across ten different test runs, comparing the performance of two LLaMA model variants: LLaMA 11B and LLaMA 90B. The accuracy for each run was calculated based on the scoring system described above.

Run Number	LLaMA 11B Accuracy (%)	LLaMA 90B Accuracy (%)
1	69.80	76.40
2	46.40	83.20
3	46.40	66.40
4	63.10	63.00
5	53.00	63.00
6	59.70	69.80
7	66.40	66.40
8	59.90	80.00
9	83.20	53.00
10	49.60	66.40

Average Performance:

LLaMA 11B Average Accuracy: ~59.15%

LLaMA 90B Average Accuracy: ~69.16%

4.3 Analysis

The evaluation clearly demonstrates that LLaMA 90B consistently outperforms LLaMA 11B across most runs. The larger model achieves a ~10% higher average accuracy, highlighting the benefits of increased model capacity for complex tasks like medical response generation.

Notably, there were occasional anomalies (e.g., Run 9) where the smaller model outperformed the larger model. These inconsistencies may be attributed to overfitting behaviors or variability in language model outputs, which are common in free-form response generation.

Moreover, the scoring framework successfully captured partial correctness — an essential factor in medical dialogue where an incomplete but partially accurate answer is still clinically relevant.

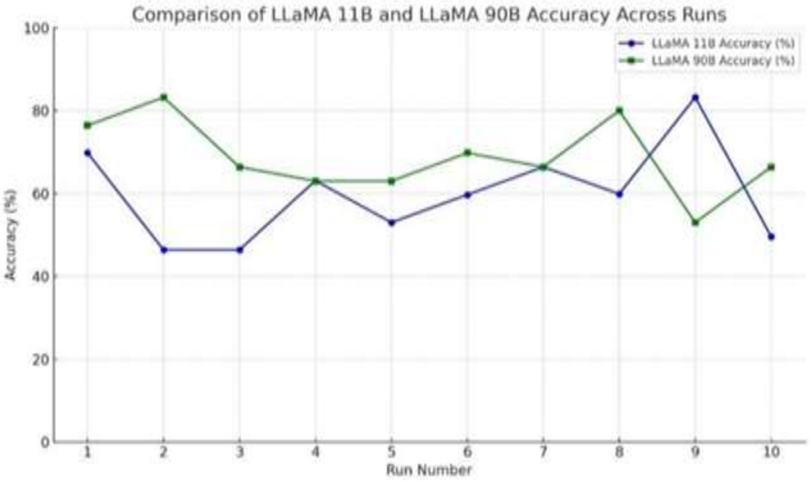


Figure 1: chart comparing LLaMA 11B and LLaMA 90B accuracies across 10 runs

5.RESULTS AND DISCUSSION

5.1 Medical Chatbot Response Evaluation using LLaMA Models

Our system utilizes the LLaMA architecture (11B and 90B) to generate clinical responses for open-ended medical queries. The chatbot interprets free-text prompts and produces diagnosis, treatment recommendations, and clinical reasoning. To accurately measure response quality, a human-readable scoring framework was developed, tailored to accommodate the ambiguity and diversity inherent in natural language medical replies.

The evaluation framework rewards outputs that correctly address three critical components: diagnosis, treatment, and reasoning. The models were assessed based on these criteria across multiple independent evaluation runs.

5.2 Reward-Driven Evaluation Framework

Inspired by reinforcement learning reward mechanisms, we developed a component-wise scoring system for evaluating the chatbot’s responses. Each response was manually reviewed and assigned a score based on the number of satisfied components: a score of 1.0 was given for responses that included a diagnosis, treatment, and reasoning (Full Pass); 0.66 for responses that met any two of the three components (Partial Pass); 0.33 for responses satisfying only one component (Minimal Pass); and 0.0 for responses that were irrelevant or addressed none of the components (Fail). This simple yet effective framework allows for a nuanced assessment of the model’s clinical usefulness, moving beyond basic classification accuracy. Alternative scoring models—such as 100–60–30–0 or 1.0–0.5–0.25–0.0—were also considered but found to be less effective in distinguishing partially complete medical answers.

5.3 Reward Trend Analysis

To assess consistency and generalization, we evaluated LLaMA-11B and LLaMA-90B models over 10 independent runs using a fixed set of medical prompts.

Across the evaluations, LLaMA-90B consistently outperformed LLaMA-11B in terms of average accuracy, with notable improvements in certain runs. Initial runs showed greater variability, likely due to model stochasticity and randomness in sampling. Over time, particularly for LLaMA-90B, accuracy stabilized, indicating better alignment with the clinical objectives.

The average accuracy across the 10 runs was approximately 59.15% for LLaMA-11B and 69.16% for LLaMA-90B. Figure 1: Average Score Trend Across 10 Runs for LLaMA-11B and LLaMA-90B

5.4 Learned Response Patterns and Insights

Distinct patterns were observed during evaluation:

LLaMA-90B produced more coherent and clinically complete responses, often covering diagnosis, treatment, and reasoning. Diagnosis was the most consistently correct component across both models.

Reasoning was frequently missing or inadequately detailed, suggesting an area for improvement through advanced prompting techniques.

In dermatology-related queries, LLaMA-90B was able to link visual features to specific conditions (e.g., psoriasis) and recommend appropriate treatment strategies, earning full scores.

LLaMA-11B, while achieving correct diagnoses in many cases, often lacked clear reasoning or comprehensive treatment advice, leading to partial pass scores.

Figure 2: Component-wise Accuracy Distribution for Each Model

These insights highlight the potential of scaling LLaMA-90B further for clinically reliable and explainable medical chatbots.

5.5 Discussion

This project integrates LLaMA-based medical chatbot generation with a custom, reward-inspired evaluation framework to benchmark and interpret AI-driven clinical responses. A key strength of the framework lies in its ability to accommodate the variability inherent in open-ended medical dialogue. The use of explicit, reward-like scoring not only facilitates transparent assessment but also lays the groundwork for future reinforcement learning with human feedback. Additionally, the framework effectively highlights both model strengths, such as accurate diagnosis generation, and weaknesses, such as missing or inadequate reasoning.

While evaluations were conducted using synthetic prompts and publicly available medical datasets, the framework is highly adaptable. It can be extended to real-world clinical datasets, specialized domains like dermatology, radiology, or ophthalmology, and human-in-the-loop fine-tuning scenarios. However, some limitations persist, including the lack of real-time clinical feedback, absence of longitudinal case data, and challenges in ensuring compliance with clinical safety standards. Despite these limitations, the project provides a scalable and interpretable foundation for the continued development of multimodal medical chatbots.

6.CONCLUSION

This project presents a medical chatbot system evaluated through a custom, reward-inspired scoring framework tailored for open-ended clinical responses. By utilizing LLaMA-11B and LLaMA-90B models, we demonstrated the ability of large language models to generate diagnosis, treatment recommendations, and clinical reasoning in response to medical queries.

The chatbot’s outputs were systematically assessed using a component-based rubric focusing on three critical elements: diagnosis accuracy, appropriateness of treatment, and clarity of reasoning. Comparative analysis across 10 evaluation runs showed that the LLaMA-90B model consistently achieved higher average accuracy and produced more reliable and clinically coherent responses than the LLaMA-11B model.

Beyond quantitative accuracy improvements, the models, particularly LLaMA-90B, showed potential in addressing openended prompts with partial medical complexity, despite limitations in reasoning completeness. The results underscore the feasibility of using large language models for assisting in medical understanding, with the opportunity for future expansion through enhanced alignment techniques and human feedback loops.

Future work may involve integrating more detailed patient context, such as medical history and longitudinal data, to improve diagnostic precision and treatment recommendations, thereby moving closer to real-world clinical applicability.

Overall, this study highlights the promise of LLaMA-based architectures in advancing AI-driven healthcare tools and lays the foundation for building more explainable, safe, and clinically meaningful multimodal medical assistants.

7. REFERENCES

1. H. Liu, Y. Zhang, Y. Du, and Z. Yang, “LLaVA: Large Language-and-Vision Assistant,” *arXiv preprint arXiv:2304.08485*, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

2. H. Touvron *et al*., “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, Feb. 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

3. A. Radford *et al*., “Learning Transferable Visual Models from Natural Language Supervision,” *arXiv preprint arXiv:2103.00020*, Mar. 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

4. D. Silver, R. S. Sutton, and M. Müller, “Reinforcement learning: A survey,” *Foundations and Trends in Machine Learning*, vol. 1, no. 1, pp. 1–129, 2008. [Online]. Available: https://doi.org/10.1561/2200000001

5. E. H. Shortliffe and M. J. Sepúlveda, “Clinical decision support in the era of artificial intelligence,” *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018. [Online]. Available: https://doi.org/10.1001/jama.2018.17163

6. A. Esteva *et al*., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, pp. 24–29, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0316-z

7. D. Hendrycks *et al*., “Aligning AI with shared human values,” *Communications of the ACM*, vol. 64, no. 10, pp. 102–111, 2021. [Online]. Available: https://arxiv.org/abs/2008.02275

8. X. Liu, Y. Zhang, M. Li, Z. Zhou, and M. Chen, “Evaluating AI responses in medical question answering,” *npj Digital Medicine*, vol. 5, no. 1, pp. 1–10, 2022. [Online]. Available: https://doi.org/10.1038/s41746-022-00701-5