Vol.9 Issue 1 Jan 2026
Multimedia Research
Received 29 October, Revised 26 November, Accepted 21 January

Evaluation of a Multimodal Custom Finetuned LLM for Virtual Healthcare Consultations

Pranav Upadhyaya
University of Mysore
pranavupadhyaya52@gmail.com

Abstract: We built a modular and privacy-focused prototype of a multimodal virtual medical assistant that uses retrieval-augmented generation (RAG) to improve healthcare consultations. The motivation behind this system is to bridge the gap between traditional telemedicine and intelligent diagnostic support by enabling AI-driven consultations that are context-aware, multimodal, and privacy-preserving. The system runs a locally deployed LLaMA 3.2 (11B) model with 4-bit quantization, keeping it lightweight yet efficient. It can process both text and images, and has been fine-tuned on 50,000 image label pairs from the MedTrinity dataset, which includes a wide range of medical images and descriptions. This fine-tuning improves the model's ability to answer multimodal medical questions. To enhance interpretability, the model's outputs are supported by transparent reasoning traces that indicate whether the response is derived from visual understanding, textual retrieval, or both.The assistant supports text, image, and speech inputs. Speech is transcribed using the AssemblyAI transcription API. For RAG, we use ChromaDB to store and retrieve medical documents from the MedQuAD dataset, which includes about 41,000 medicine-related question answer pairs. This integration enables the system to fetch domain-relevant evidence dynamically, helping users verify the medical reliability of generated responses. We evaluate our fine-tuned model against the base LLaMA 3.2 model and the responses are judged using OpenAI's GPT-4.1 as an evaluator. Performance is measured on the MMMU benchmark, focusing on three medical domains:1) Basic medical science, 2) Clinical medicine, 3) Diagnostic and laboratory medicine. Each model variant (with and without RAG) was tested on 30 questions per domain, and evaluated under strict and non-strict scoring criteria. The evaluation reveals that fine-tuning significantly enhances answer relevance and domain fluency, while RAG contributes variably depending on retrieval quality, underscoring the need for domain-specific curation in medical AI systems.

Keywords: Multimodal, Retrieval-Augmented Generation (RAG), ChromaDB, LLaMA 3.2 11B, 4-bit Quantization, GPT-4.1, Virtual Healthcare Consultations, AI in Medicine

Nomenclature

Abbreviation	Description
AI	Artificial Intelligence
ANOVA	Analysis of Variance
FT	Fine-Tuned Model
FT+RAG	Fine-Tuned Model with Retrieval-Augmented Generation
LLM	Large Language Model
RAG	Retrieval-Augmented Generation

1. Introduction

With the rapid rise of online medical consultation platforms such as Practo, India has seen a growing demand for accessible, affordable, and high-quality virtual healthcare services. These digital health systems have made it easier for patients in cities to reach qualified doctors from the comfort of their homes. However, the same accessibility is not evenly distributed across India, especially in rural and semi-urban regions, where healthcare infrastructure remains limited. The motivation behind this research arises from the urgent need to develop equitable, intelligent healthcare solutions that can extend beyond metropolitan areas and operate under real-world resource constraints. While telemedicine adoption is growing, most existing platforms primarily focus on video consultations and appointment systems rather than on intelligent, context-aware medical reasoning or diagnosis support [9][10][11][20].

Nearly 65 - 70% of India's population resides in rural areas, yet these regions face a severe shortage of medical professionals and diagnostic facilities. Many primary health centres lack specialist doctors, and patients are often required to travel long distances for even basic consultations. Connectivity issues,

https://doi.org/10.46253/j.mr.v9i1.a4

language barriers, and limited digital literacy further complicate the adoption of telemedicine services. As a result, timely diagnosis and treatment are frequently delayed particularly for chronic conditions and elderly patients who form a growing segment of the population. These challenges highlight the limitations of existing telehealth infrastructures, which often rely on cloud-dependent AI models and are not optimized for offline or privacy-sensitive deployments [15][16][17]. Therefore, researchers face the dual challenge of ensuring both accessibility and data security while maintaining diagnostic accuracy in low-resource settings.

This imbalance between urban and rural healthcare delivery highlights a pressing national concern: how to make telemedicine scalable, intelligent, and contextually relevant to diverse patient populations while maintaining privacy and clinical reliability [1]. To address this challenge, we propose a modular, privacy-conscious virtual medical assistant powered by RAG [12]. The system integrates multimodal understanding text, image, and speech to simulate a realistic medical consultation experience. By leveraging locally deployed AI models and curated medical datasets, it aims to reduce dependence on continuous cloud connectivity, assist clinicians in low-resource environments, and extend reliable teleconsultation support to underserved communities across India [10][15][16][20]. The key contributions of this study are summarized as follows:

- Implementation of a locally deployable multimodal LLaMA 3.2 (11B) model optimized using 4-bit quantization for efficiency.
- Fine-tuning with 50,000 image–text pairs from the MedTrinity dataset to enhance medical image reasoning.
- Integration of ChromaDB-based RAG using the MedQuAD dataset for contextual medical knowledge retrieval.
- Performance evaluation across three medical domains—basic medical science, clinical medicine, and diagnostic and laboratory medicine—using GPT-4.1 as an automated evaluator.

We present a novel proof of concept in which artificial intelligence can play a role in the MedTech sector, especially in connecting patients to the right doctor or providing a brief overview of their X-ray/CT-Scan images. We trained the popular lightweight multimodal LLM Llama 3.2-11 B. This model has been trained [3] with 50,000 medical image text pairs from the Med Trinity dataset [4] and uses context from a local vector embedding database (here, we use Chroma DB vector embeddings) consisting of 41,000 medical science-related question answer pairs [2]. Experimental evaluation demonstrates that fine-tuning improves response relevance and domain fluency compared to the base model while maintaining full data privacy through local inference. In the future, we plan to build an agent that escalates the patient's query to a doctor for manual review (a human-in-the-loop agentic inference), which can be responded to by the doctor on the basis of the chat log with the LLM model. This LLM model can also be run locally without the internet. A system with at least 8 GB of RAM and an NVIDIA-based GPU is needed. Inference without a GPU is currently not possible, as the Python libraries PyTorch and Unsloth [3], which are essential in running the model, require an NVIDIA CUDA inference to run.

The remainder of this paper is organized as follows: Section 2 presents the literature review and related works, Section 3 discusses the methodology and statistical analysis framework, Section 4 provides the results and discussion, and Section 5 concludes with key findings and directions for future work.

2. Literature Review

2.1 Related Work

Recent advances in MLLMs and RAG have opened new directions for AI-assisted healthcare. These models combine visual, textual, and sometimes speech modalities to interpret complex clinical data, enabling applications in diagnostics, radiology reporting, and clinical decision support. Early works such as Med-PaLM 2 [21] and Med-Flamingo [22] demonstrated that aligning large language models to medical question-answering tasks can yield near-expert-level performance, while later systems such as LLaVA-Med and MMed-RAG [14] extended these methods to visual reasoning and multimodal retrieval. Despite their strong performance, most of these systems rely on cloud-hosted models with high computational requirements and potential risks to patient data privacy, limiting their applicability in low-resource or privacy-regulated settings. Table I summarizes key representative systems, their datasets, and findings, providing context for how our prototype differs by emphasizing privacy-conscious local deployment, 4-bit quantization, and fine-tuning versus RAG trade-offs in medical consultation scenarios. Our system builds upon these foundational efforts by integrating both multimodal understanding and retrieval capabilities within a lightweight, locally deployable environment, thereby bridging the gap between research-grade performance and practical healthcare implementation. Unlike prior systems, our approach explicitly

balances fine-tuning and retrieval in a privacy-first local deployment setting, addressing real-world feasibility in healthcare environments.

Table 1. *Representative Multimodal and RAG-Based Medical Assistant Systems*

Year	Authors	System / Dataset	Key Findings
2023	Singhal et al.	Med-PaLM / Med-PaLM 2 (MedQA, USMLE)	First LLMs to surpass USMLE pass mark; domain alignment improved factual accuracy in clinical Q&A.
2023	Moor et al.	Med-Flamingo (image–text pairs, VQA)	Few-shot multimodal reasoning; improved clinician ratings and rationale generation.
2023	Li et al.	LLaVA-Med (PubMed Central figures + captions)	Biomedical vision–language assistant trained efficiently; strong performance on open VQA tasks.
2023	Bannur et al.	BioViL-T (chest X-rays + reports)	Temporal vision–language pretraining improved radiology report understanding.
2024	Xia et al.	MMed-RAG (radiology/ophthalmology/pathology)	Domain-aware retrieval + adaptive context; 43.8 % factual-accuracy improvement in Med-LVLMs.

3. Methodology

3.1 Statistical Analysis of the Evaluation Data

To rigorously assess the performance of our multimodal LLM prototypes, we adopted an LLM-as-a-judge framework to perform both qualitative and quantitative evaluations.

LLM-as-a-judge evaluation was adopted in place of human expert review primarily due to its scalability, cost-efficiency, and consistency across large evaluation batches. While domain experts in medical science provide high-quality subjective feedback, manual evaluation of 360 responses (90 questions × 4 model variants) would be time-intensive and prone to inter-rater variability. Using GPT-4.1 as an automated evaluator ensured standardized scoring criteria, objective comparison across models, and rapid reproducibility of results, aligning with recent best practices in multimodal LLM benchmarking [7].

In this evaluation, we independently assessed the system using two distinct frameworks—LangChain's evaluation and DeepEval's answer relevancy. Both employed GPT-4.1 as the judging LLM, but with differing strictness levels, allowing a more comprehensive comparison of model behavior across evaluation settings.

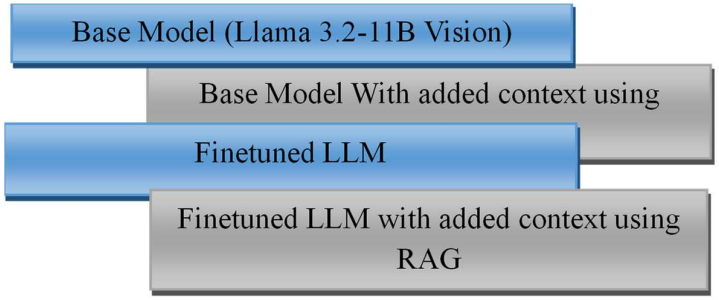

Fig 1. *Overview of evaluated LLaMA 3.2–11B model configurations*

The evaluation pipeline was designed to measure the accuracy, relevance, and contextual consistency of generated medical responses across four model variants:

- **Base Model (LLaMA 3.2 11B Vision)**
- **Base + RAG** (Base model enhanced with retrieval context)
- **Fine-Tuned (FT)** model (trained on MedTrinity 50 k pairs)
- **Fine-Tuned + RAG (FT + RAG)**

Thirty questions were selected from each of the three domains in the MMMU benchmark—Basic Medical Science, Clinical Medicine, and Diagnostic & Laboratory Medicine—making a total of 90 questions per evaluation framework. Each model variant produced responses to every question, and these were evaluated under the following two frameworks.

3.1.1 Evaluation Design

We selected 30 representative questions from each of the three core domains in the MMMU dataset basic medical science, clinical medicine, and diagnostic & laboratory medicine resulting in a total of 90 questions per evaluation framework. Each model variant generated an answer for every question, and the outputs were independently evaluated under two complementary judging frameworks:

- **LangChain Evaluation Framework**: uses GPT-4.1 as a grading model under a strict evaluation mode, assigning binary scores (1 = relevant, 0 = irrelevant). This setup emphasizes precision and factual alignment, penalizing partially correct answers. LangChain's scoring design corresponds to a "strict" metric, meaning even slightly incomplete or contextually inaccurate answers are scored as 0. The final score is calculated as the mean of all question-level scores,
- **DeepEval Answer Relevancy Framework**: also employs GPT-4.1 as the scoring agent but applies a non-strict metric, allowing partial credit for answers that are semantically close or contextually relevant. This framework focuses on semantic coherence and pragmatic relevance, making it more lenient than LangChain's strict evaluation.

Both frameworks thus offer complementary insights—LangChain quantifies factual correctness, while DeepEval measures contextual relevance and reasoning consistency.
For both frameworks, the mean evaluation score per model variant was computed as:

$$Mean\,Score = \frac{\sum LLM\ Scores}{N} \tag{1}$$

where $N = 30 \times 3 = 90$ total questions per evaluation type.

3.1.2 Statistical Testing Framework

To identify whether observed score differences were statistically meaningful rather than due to chance, we applied non-parametric hypothesis testing methods suitable for ordinal or binary data.

3.1.2.1 Friedman's Test

We first performed Friedman's test, a non-parametric alternative to the one-way repeated-measures ANOVA. It evaluates whether there are systematic rank differences among multiple related samples in this case, the four model variants (Base, Base+RAG, FT, FT+RAG) evaluated on identical question sets.
This test was chosen because:

- The evaluation scores are not normally distributed (binary or bounded between 0 – 1).
- The same questions were used for each model, satisfying the repeated-measures condition

The Friedman test statistic X_F^2 and associated p-value were computed separately for LangChain and DeepEval frameworks using the following formula:

$$F = \left[\frac{12}{N * k * (k+1)} * \sum R^2 - 3 * N * (k+1) \right] \tag{2}$$

A significance threshold of **p < 0.05** was used to determine overall differences among models.

3.1.2.2 Post-hoc Wilcoxon Signed-Rank Tests

Following Friedman's global test, we performed pairwise Wilcoxon signed-rank tests to analyze differences between every pair of model variants. The Wilcoxon test measures whether the median of paired differences between two related samples significantly deviates from zero.

To control for Type I error due to multiple comparisons, a Bonferroni correction was applied, yielding an adjusted significance threshold of

$$\alpha_{corrected} = \frac{0.05}{6} = 0.0083 \tag{3}$$

since there are six possible pairwise comparisons.
The test statistic (W) was computed as:

$$W = \sum \mathrm{sgn}(x_{1i} - x_{2i}) R_i \tag{4}$$

where R_i is the rank of the absolute difference between pairs.

This post-hoc analysis complements the Friedman test by revealing which pairs (if any) differ significantly.

3.1.2.3 Percentage Shift Analysis

Beyond significance testing, we calculated percentage shift values to capture directional performance trends between model variants. This is expressed as:

$$Percentage\,Shift = \frac{\mu_2 - \mu_1}{\mu_1} \times 100 \tag{5}$$

where μ_1 and μ_2 are the mean scores of two compared models.This allows visualization of improvements or declines (positive or negative shifts) even when statistical significance is absent.

Heatmaps were generated to visualize pairwise percentage shifts, facilitating intuitive comparison between the LangChain and DeepEval frameworks.

3.1.3 Results Overview

Friedman tests showed no statistically significant global differences across the four model variants in either framework ($p > 0.05$). This means all four models performed at similar levels.Wilcoxon tests, with Bonferroni correction, similarly found no significant pairwise differences.
However, the percentage shift analysis provided nuanced insights:

- **Fine-Tuning (FT)** improved relevance modestly (+2.57 % in DeepEval, +13.91 % in LangChain). Fine-tuning made the model more accurate and consistent.
- **Adding RAG** to the fine-tuned model often reduced performance (−2.73 % in DeepEval, −25.69 % in LangChain), suggesting that RAG introduced retrieval noise or mismatched context. The RAG data may not have matched the question topics well.
- **Base+RAG** showed mixed behavior, improving slightly in LangChain (+3.24 %) but dropping in DeepEval (−5.78 %). Retrieval helped in some strict cases but hurt broader relevance.

These findings indicate that fine-tuning contributes more to answer accuracy than retrieval augmentation, especially when RAG sources are not perfectly aligned with the question domain

Table 2. *Summary of Analytical Approach*

Step	Purpose	Method	Output
Data Scoring	Assign answer quality scores	LangChain (strict) & DeepEval (non-strict)	Mean relevance scores
Global Difference Test	Assess overall variation among models	Friedman's test	χ^2F and p value
Pairwise Comparison	Identify specific model differences	Wilcoxon signed-rank + Bonferroni	Pairwise p values
Trend Analysis	Measure direction of change	Percentage shift analysis	% improvement / decline & heatmaps

4. Results and Discussion

The statistical evaluation compared four model variants—Base, Base + RAG, Fine-Tuned (FT), and Fine-Tuned + RAG—across two evaluation frameworks, LangChain (strict) and DeepEval (non-strict).

4.1 Friedman Test

Table 3 presents the Friedman test outcomes, which showed no statistically significant global differences ($p > 0.05$) among the models in either framework. This suggests that all configurations performed comparably in overall relevance scores, with no single model achieving a clear quantitative advantage.

Table 3. *Friedman test results comparing the four model configurations under both LangChain (strict) and DeepEval (non-strict) frameworks. Both tests produced p-values greater than 0.05, indicating no statistically significant global difference among models.*

Evaluation Type	Result	Discussion
DeepEval Answer Relevancy	F = 5.8000, p = 0.1218	Since p > 0.05→ No significant differences.
Langchain Evaluation	F = 2.2800, p = 0.5164	Since p > 0.05→ No significant differences.

4.2 Wilcoxon Signed Rank Test

Tables 4 and 5 display the Wilcoxon signed-rank test results for the DeepEval and LangChain frameworks, respectively. After Bonferroni correction ($\alpha = 0.0083$), no pairwise comparison reached statistical significance. This indicates that although some models showed directional performance changes, the variations were not strong enough to be statistically confirmed within the sample size used (N = 90).

4.2.2 Wilcoxon Signed Rank Test for DeepEval Answer Relevancy

Table 4. *Pairwise Wilcoxon signed-rank test results with Bonferroni-adjusted α = 0.0083 for DeepEval evaluation. No pairwise comparison showed a statistically significant difference, confirming the Friedman test's global results*

Categories Pairwise Evaluation	Result	Discussion (Bonferroni-corrected α = 0.0083)
Base vs Base+RAG	0.2500	not significant
Base vs FT	0.7500	not significant
Base vs FT+RAG	1.0000	not significant
Base+RAG vs FT	0.2500	not significant
Base+RAG vs FT+RAG	0.2500	not significant
FT vs FT+RAG	0.7500	not significant

4.2.3 Wilcoxon signed rank test for chain evaluation

Table 5. *Pairwise Wilcoxon signed-rank test results with Bonferroni-adjusted α = 0.0083. No pairwise comparison showed a statistically significant difference, confirming the Friedman test's global results.*

Categories Pairwise Evaluation	Result	Discussion (Bonferroni-corrected α = 0.0083)
Base vs Base+RAG	1.0000	not significant
Base vs FT	1.0000	not significant
Base vs FT+RAG	0.7500	not significant
Base+RAG vs FT	1.0000	not significant
Base+RAG vs FT+RAG	0.5000	not significant
FT vs FT+RAG	0.5000	not significant

4.3 Percentage Shift Values

Table 6 summarizes the percentage-shift analysis, offering a more nuanced view of model behavior. Fine-tuning consistently improved answer quality in both frameworks (+2.57 % in DeepEval, +13.91 % in LangChain), reflecting better domain adaptation from the MedTrinity dataset. Conversely, introducing RAG caused inconsistent or negative shifts (−2.73 % in DeepEval, −25.69 % in LangChain), suggesting that retrieved context sometimes introduced off-topic or redundant information.

Table 6. *Comparative visualization of mean score percentage shifts across the two evaluation frameworks. The Fine-Tuned model shows consistent improvement, while the addition of RAG introduces varying effects depending on the framework.*

Category	DeepEval Answer Relevancy	Langchain Evaluation
Base → Base+RAG	-5.78% change	3.24% change
Base → FT	2.57%	13.91% change
Base → FT+RAG	-0.24% change	-15.35% change
FT →Base+RAG	8.86% change	10.33% change
Base+RAG → FT+RAG	5.88% change	-18.01% change
FT →FT+RAG	-2.73% change	-25.69% change

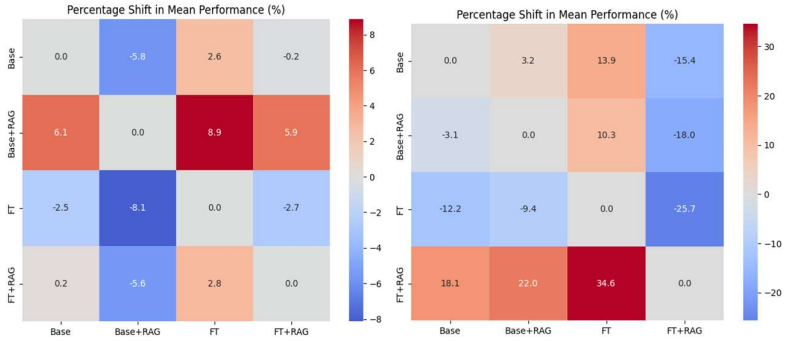

Fig 2. *Workflow of the statistical analysis pipeline, including evaluation score computation, Friedman's nonparametric global test, post-hoc Wilcoxon signed-rank testing with Bonferroni correction, and visualization of percentage shift heatmaps.*

We compared four model configurations (Base, Base+RAG, FT, FT+RAG) using both percentage-shift metrics and nonparametric tests. This dual analysis helped capture not only statistical significance but also directional performance trends between model types.

The percentage shift analysis provided a directional view of model behaviour across the two evaluation frameworks. It offered a more intuitive understanding of how each modification—fine-tuning or retrieval—impacted model accuracy and contextual relevance.

- Under the DeepEval framework, which emphasizes semantic relevance, fine-tuning led to a modest +2.57% improvement compared to the base model. However, adding RAG to the fine-tuned model caused a −2.73% decline, suggesting that retrieval augmentation may have introduced irrelevant or mismatched content from the database. This indicates that while fine-tuning enhances the model's internal medical understanding, RAG can sometimes inject noisy external context that reduces precision.
- In contrast, within the LangChain framework (which applies stricter binary relevance scoring), fine-tuning achieved a much larger +13.91% performance gain, while the addition of RAG resulted in a sharp −25.69% performance drop. The strong improvement through fine-tuning under strict scoring highlights its effectiveness in producing concise, factually accurate response

- The Base + RAG configuration performed inconsistently—showing a −5.78% drop in DeepEval but a +3.24% gain in LangChain—further indicating that the benefit of retrieval augmentation is highly dependent on the evaluation metric and data alignment. These variations suggest that retrieval mechanisms must be carefully aligned with the task domain to avoid undermining model reliability.

4.4 Interpretation of Fine-Tuning Performance

Fine-tuning consistently improved model outputs across both frameworks because it directly optimized the model on domain-specific multimodal data 50,000 image–text pairs from the MedTrinity[4] dataset. Exposure to this large-scale, diverse dataset helped the model internalize complex visual–textual associations relevant to real-world diagnostic reasoning.This targeted exposure allowed the model to better associate visual medical features (e.g., X-ray, MRI, or pathology images) with their corresponding textual findings and terminology. As a result, the fine-tuned system was able to produce contextually richer and clinically coherent responses compared to the base version.

As a result, the fine-tuned model developed stronger contextual grounding and domain fluency, leading to more accurate and coherent answers during medical Q&A. It also demonstrated improved recall of domain-specific terminology and diagnostic language structures.In other words, fine-tuning improved the model's internal representation of medical knowledge rather than relying solely on external retrievals.

4.5 Why RAG Underperformed

By contrast, the RAG mechanism did not consistently enhance performance and, in some cases, degraded response quality. Several factors explain this behaviour:
Although RAG is conceptually valuable for grounding responses in external knowledge, its performance depends heavily on retrieval accuracy and contextual alignment with the query domain.

- **Domain Misalignment of Retrieved Contexts:** The ChromaDB index was built using the MedQuAD dataset, which contains general medical question–answer pairs. However, the questions in the MMMU [8] benchmark are often more specialized or diagnostic in nature. This mismatch likely caused the RAG module to retrieve semantically related but clinically irrelevant passages [19], diluting the precision of the generated response. Inaccurate retrieval led to information drift, where the model's response deviated from the intended clinical focus.
- **Context Overload and Attention Saturation**: When excessive retrieval data is appended to the prompt, the LLM may assign attention to less relevant tokens, leading to noise in the reasoning chain. This effect is amplified in strict evaluation frameworks like LangChain, where partially correct or irrelevant responses are penalized. The surplus context fragments the model's attention, reducing its ability to prioritize key diagnostic cues.
- **Limited Retrieval Precision**: Since RAG relies on semantic embeddings rather than structured reasoning, it may retrieve conceptually similar but contextually mismatched answers (e.g., mixing conditions with similar symptoms). This problem is particularly evident in multimodal medical questions, where subtle image–text correlations matter [13]. Such mismatched retrievals undermine factual consistency and reduce interpretability, especially in clinical scenarios demanding high specificity.

Consequently, the Fine-Tuned model without RAG emerged as the most balanced configuration, achieving steady improvements across both evaluation schemes. It maintained domain accuracy while avoiding retrieval-induced noise.These results highlight that, in specialized domains such as healthcare, data alignment and knowledge consistency are often more valuable than adding external retrieval components.

4.6 Qualitative Examples and Observations

To complement the quantitative analysis, Table II presents sample question–answer pairs from the MMMU evaluation set. These illustrate the characteristic behaviour of each model variant and clarify why fine-tuning improved relevance whereas RAG sometimes reduced coherence. The examples highlight the model's reasoning quality, clinical correctness, and tendency to hallucinate under mismatched retrieval contexts.

Table 7. *Example qualitative comparisons showing improvement through fine-tuning and degradation under misaligned retrieval.*

Query ID / Domain	User Prompt (Condensed)	Model Variant	Generated Response (Excerpt)	Reviewer Comment / Observation
Q-17 / Clinical Medicine[8]	"A 55-year-old diabetic presents with non-healing ulcer on the foot. What initial imaging is recommended?"	Base	"The doctor should do MRI for diagnosis."	Partially correct but lacks reasoning; inconsistent with clinical guidelines.
		Base + RAG	"MRI or CT scan may be needed depending on soft-tissue swelling (per MedQuAD ref #134)"[2]	Retrieval adds factual context but mixes unrelated foot-trauma content. Minor hallucination.
		FT	"Start with plain X-ray to assess bone involvement; MRI only if osteomyelitis suspected."	Clinically accurate, concise, and context-aware; aligns with domain knowledge from MedTrinity pairs.
		FT + RAG	"X-ray should be done, and the patient should consult orthopaedics for trauma management (per MedQuAD)" [2]	Retrieval distracts focus toward trauma rather than diabetic ulcer. Accuracy drop.
Q-32 / Diagnostic Lab Medicine[8]	"Interpret a chest X-ray showing right lower-lobe opacity with air bronchogram."	Base	"Possible pneumonia."	Too brief; lacks modality reasoning.
		FT	"Consolidation of right lower lobe suggestive of bacterial pneumonia; correlate with fever and sputum culture."	Richer diagnostic phrasing; demonstrates visual and text understanding learned via fine-tuning.
		FT + RAG	"Opacity indicates atelectasis or pneumonia; refer to MedQuAD case #512" [2]	Adds citation but hedges diagnosis; retrieval noise lowers confidence.
Q-58 / Basic Medical Science[8]	"Differentiate Gram-positive and Gram-negative cell walls."	Base	"They differ by stain color."	Correct but superficial.
		FT	"Gram-positive have thick peptidoglycan and teichoic acids; Gram-negative have outer membrane with LPS."	Accurate and complete. Evidence of improved factual recall.
		FT + RAG	"Gram-positive appear purple; Gram-negative pink per MedQuAD topic #220"[2]	Retrieval reinforces basic color detail but omits structural distinctions.

4.7 Statistical Observations

Despite these observable trends, nonparametric hypothesis testing revealed no statistically significant differences across the models:
This suggests that the measured performance differences, while directionally consistent, were not strong enough to reach statistical significance under the chosen thresholds.

- The Friedman test produced p values of 0.1218 (DeepEval [7]) and 0.5164 (LangChain), both above the 0.05 threshold, indicating no global differences among the four model types. Hence, the observed gains from fine-tuning are indicative rather than statistically conclusive.
- Similarly, Wilcoxon signed-rank tests with Bonferroni-corrected $\alpha = 0.0083$ found no significant pairwise differences across any comparison. This consistency across frameworks strengthens the reliability of the statistical findings.
- These findings suggest that, while fine-tuning directionally improves performance, the differences were not statistically robust under the current dataset size ($N = 90$). Increasing sample size or expanding domain diversity could yield more definitive evidence of model improvement.

5. Future Works

Future work should incorporate larger sample sizes and more diverse medical question sets to validate these trends. Broader evaluation across multiple medical domains would help test the model's adaptability and robustness. Additionally, retraining or filtering the retrieval corpus for domain-specific precision may enhance the effectiveness of RAG without introducing noise. Future studies could also explore adaptive retrieval or hybrid fine-tuning methods to improve contextual alignment. Further optimization for offline deployment and integration with clinician feedback systems could make the model more practical for real-world medical use.

6. Conclusion

This study presented a modular, privacy-conscious prototype of a multimodal virtual medical assistant that integrates fine-tuned LLaMA 3.2 (11 B) [5] with RAG. The system shows how local AI deployment can balance accuracy, privacy, and efficiency. Evaluation using LLM-as-a-judge frameworks, revealed that fine-tuning improves medical answer relevance while RAG requires better alignment to avoid contextual noise. This highlights that high-quality domain data is more valuable than additional retrieval complexity. The 4-bit quantization [6] and on-device inference enable efficient, confidential operation even on modest GPUs, supporting India's data localization policies. Local inference also ensures usability in low-connectivity healthcare environments. Future work will expand fine-tuning to broader multimodal datasets (pathology, radiology, dermatology [19][18]) and apply federated techniques for privacy-preserving collaboration. Integration with clinician feedback loops could enhance trust and real-world adoption. Overall, this research contributes a cost-efficient, ethically aware AI framework, combining technical efficiency with privacy-by-design to advance responsible and equitable healthcare innovation across both urban and rural communities [1].

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

Declaration of generative AI and AI-assisted technologies in the writing process: The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Arvind Kasthuri (2018). Challenges to Healthcare in India - The Five As. https://pmc.ncbi.nlm.nih.gov/articles/PMC6166510/

[2] Asma Ben Abacha and Dina Demner Fushman (2019). A Question-Entailment Approach to Question Answering. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3119-4

[3] Daniel Han, Michael Han, and Unsloth team (2023). Unsloth. http://github.com/unslothai/unsloth

[4] Yunfei Xie and Ce Zhou and Lang Gao and Juncheng Wu and Xianhang Li and Hong-Yu Zhou and Sheng Liu and Lei Xing and James Zou and Cihang Xie and Yuyin Zhou. MedTrinity-25 M. A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine https://arxiv.org/abs/2408.02900

[5] Touvron, H., et al. (2024). Llama 3 Herd of models. https://arxiv.org/abs/2407.21783

[6] Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, Zettlemoyer, and Luke. QLoRA: Efficient fine-tuning of quantized llms. https://arxiv.org/abs/2305.14314

[7] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. https://arxiv.org/abs/2303.16634

[8] Xiang Yue and Yuansheng Ni and Kai Zhang and Tianyu Zheng and Ruoqi Liu and Ge Zhang and Samuel Stevens and Dongfu Jiang and Weiming Ren and Yuxuan Sun and Cong Wei and Botao Yu and Ruibin Yuan and Renliang Sun and Ming Yin and Boyuan Zheng and Zhenzhu Yang and Yibo Liu and Wenhao Huang and Huan Sun and Yu Su and Wenhu Chen. MMMU: A massive multidisciplinary multimodal understanding and Reasoning Benchmark for Expert AGI. https://arxiv.org/abs/2311.16502

[9] R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, A. Ahmed, M. A. Renault, R. Damseh, et al., "Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook," J. Med. Internet Res., vol. 26, p. e59505, 2024. PMC

[10] Y. Hu, et al., "Review: Medical Multimodal Large Language Models," ScienceDirect, 2025. ScienceDirect

[11] Zabir Ali Nazi, "Large Language Models in Healthcare and Medical Domain," Inf. (MDPI), vol. 11, no. 3, p. 57, 2024. MDPI

[12] "Enhancing medical AI with retrieval-augmented generation," PMC, 2025. PMC

[13] "Comprehensive and Practical Evaluation of Retrieval-Augmented Generation for Medical QA," arXiv, 2024. arXiv

[14] X. Zhao, S. Liu, S.-Y. Yang, C. Miao, "MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot," arXiv, 2025. arXiv

[15] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, "Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation," arXiv, 2024. arXiv

[16] L. Buess, et al., "A Scoping Review on the Potential of Generative AI in Medicine," Springer, 2025. SpringerLink

[17] "Retrieval-Augmented Generation in Biomedicine: A Survey," arXiv, 2025. arXiv

[18] P. Gupta, "Growing usage of Multimodal Large Language Models in Medicine," ScienceDirect, 2025. ScienceDirect

[19] "Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models," arXiv, 2024. arXiv
[20] H. Xiao, F. Zhou, X. Liu, T. Liu, Z. Li, X. Liu and X. Huang, "A Comprehensive Survey of Large Language Models and Multimodal Large Language Models in Medicine," *arXiv*, May 2024. arxiv.org
[21] R. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.
[22] A. Moor et al., "Med-Flamingo: a multimodal medical few-shot learner," Proc. Mach. Learn. Res., 2023.