International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Evaluation of a Multimodal Custom Finetuned LLM for Virtual Healthcare Consultations

Pranav Upadhyaya

University of Mysore

Abstract:
We present a modular, privacy-conscious prototype for multimodal agency with retrieval-augmented generation (RAG) for a virtual medical assistant in healthcare consultation. The system features a locally deployed LLaMA 3.2 11B with 4-bit quantization to keep the model small yet efficient. The model directly accepts both images and text and has been fine-tuned using 50,000 image label pairs. The image label pairs are taken from the MedTrinity dataset, which consists of a wide variety of medical-related image-text pairs. The model was fine-tuned to enhance multimodal query answering in medical contexts. Text, image, and speech inputs are all supported. Speech is transcribed via the Assembly AI transcription API. For retrieval-augmented generation, ChromaDB semantically stores indexed medical documents sourced from the MedQuAD dataset, where 41,000 medicine-related question–answer pairs are stored.
We evaluate the finetuned model by comparing it with the base model, both of which are compared with and without the support of Retrieval Augmented Generation (RAG). We assess the response via LLM as a judgement criterion via OpenAI's GPT-4.1. We use strict vs nonstrict evaluations of the model against the MMMU benchmark. For the MMMU dataset, we select the fields of basic medical science, clinical medicine, and diagnostic & laboratory medicine. Each field was evaluated with 30 questions per LLM variant with or without RAG support.

Keywords: Multimodal, Retrieval-Augmented Generation (RAG), ChromaDB, LLaMA 3.2 11B, 4-bit Quantization, GPT-4.1

1. Introduction

With the rise of online medical consultation platforms and services, such as Practo, we have seen a surge in demand for quality yet affordable healthcare solutions that connect healthcare professionals to patients across India, whether in rural or urban areas. With an increase in the aging population and the looming healthcare crisis[1], virtual consultation startups are set to play a growing role in expanding the MedTech space.
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 with 50,000 medical image text pairs from the MedTrinity dataset and uses context from a local vector embedding database (here, we use ChromaDB vector embeddings) consisting of 41,000 medical science-related question–answer pairs [2].
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, which are essential in running the model, require an NVIDIA CUDA inference to run.

2. Statistical analysis of the evaluation data

We use LLM-as-a-judge evaluation. We independently investigated the system via two different frameworks, i.e., LangChain's evaluation and DeepEval's answer relevancy. Thirty questions are selected from each field in the MMMU dataset. A score is assigned to each LLM answer for each of the following evaluation criteria:

Base Model (Llama 3.2-11B Vision)
Base Model With added context using RAG
Finetuned LLM
Finetuned LLM with added context using RAG

Fig. 1

LLM as a Judgement with Langchain

Langchain's evaluation framework uses GPT-4.1 by default for evaluation unless otherwise specified. The score assigned was either 0 for a less relevant LLM response or 1 for more relevant LLM prompts. This indicates a stricter evaluation type. The final score generated was calculated as the sum of all LLM scores divided by the total number of questions (30) per domain in the MMMU dataset. LLM scores were calculated for each of the aforementioned criteria.

LLM as a Judgement with DeepEval

DeepEval's answer relevance framework uses GPT-4.1 only for evaluation. Since the "strict" metric was not set as true, it did not assess the LLM's prompt in a strict evaluation type such as Langchain's evaluation framework. We investigate the relevancy of the LLM's response against 30 questions per field in the MMMU dataset. We infer an average score, assuming that the maximum score possible per evaluation question is 1. We obtain scores per the evaluation criteria mentioned in Fig. 1.

Statistical analysis of the evaluation data

We first perform Friedman's nonparametric test on the given samples per the criteria. Friedman's test is equivalent to a nonparametric ANOVA test, as it does not make any assumption before performing the statistical calculation.

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

(where N is the number of individuals, k is the number of conditions (measurements or samples), and R is the total rank for each of the columns of data.)

Fig. 2 – Friedman's test formulae

Post hoc, after the Friedman test, we perform the Wilcoxon signed rank test with Bonferroni correction. This indicates pairwise differences between samples and whether they are relevant enough to be considered significant.

$$W = |\sum[\mathrm{sgn}(x_{2,} - x_{1}) \cdot R]|$$

(
W = test statistic
Nr = sample size, excluding pairs where x1 = x2
sgn = sign function
x1i, x2i = corresponding ranked pairs from two distributions
Ri = rank i
)

Fig. 3 – Wilcoxon signed rank test formulas

This is an important step, as the Wilcoxon signed rank test tries to calculate the pairwise difference between pairs of individual samples, whereas Friedman's test considers all the criterion samples together. This provides further insight into the pairwise differences between each sample.

$$\frac{New\ value - Old\ value}{Old\ Value} \times 100$$

Fig. 4 – Formula for calculating percentage shift

Finally, we calculate the percentage shift between the means of individual samples and calculate the percentage shift between the mean values to obtain final insight into the difference in performance between all four criteria used in the evaluation. We also generate a heatmap of the pairwise percentage shift values to obtain final insight via a visual representation of the values.

System Specifications during Testing

Testing Platform – Google Colab: Python 3 Google Compute Engine backend (GPU)
System RAM – 53 GB
GPU RAM – 22.25 GB
Disk space – 112.6 GB
Libraries used – Unsloth[3], Transformers[Huggingface], Datasets[Huggingface], Langchain, ChromaDB, Pillow[Python Imaging Library], PyTorch.

3. Results and Discussion

Friedman test

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.

Fig. 5 – Friedman test results by evaluation type.

Wilcoxon signed rank test

Wilcoxon signed rank test for DeepEval answer relevancy

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

Fig. 6 - Wilcoxon signed rank test (1)

Wilcoxon signed rank test for chain evaluation

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

Fig. 7 - Wilcoxon signed rank test (2)

Percentage Shift Values

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. 8 – Percentage shift values

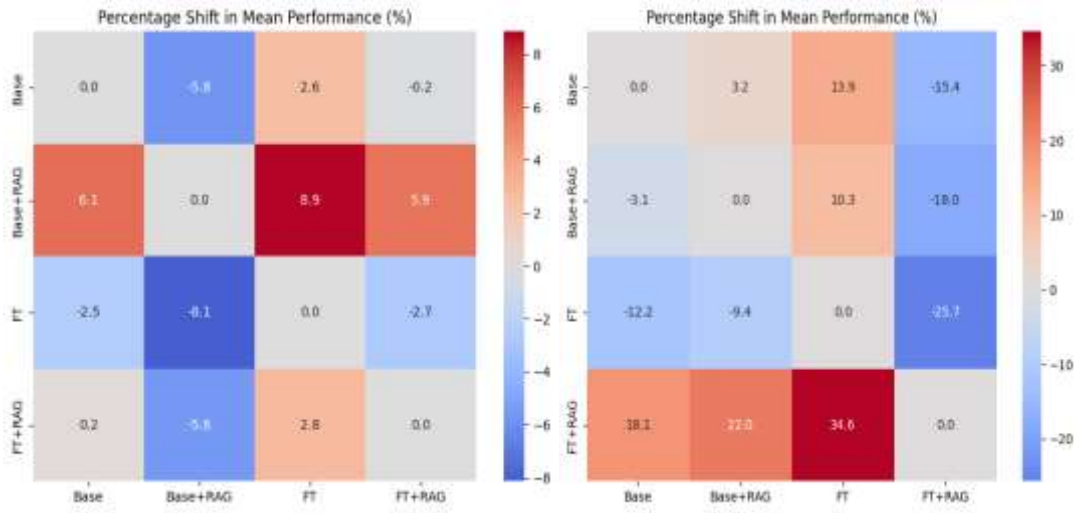

Fig. 9 - Comparison of percentage shift heatmaps between the Langchain and DeepEval evaluations.

To assess the performance differences between categories - Base, Base+RAG, FT, and FT+RAG. We conducted both metric-based percentage shift analysis and formal hypothesis testing. The percentage shift results revealed nuanced trends across the two evaluation frameworks. For example, DeepEval (focused on answer relevancy), the fine-tuned model led to a modest +2.57% improvement, whereas adding RAG to the fine-tuned model resulted in a −2.73% decline. Under Langchain evaluation, fine-tuning yielded a +13.91% improvement, whereas the addition of RAG caused a substantial −25.69% drop, suggesting that RAG may introduce noise or misalignment when combined with fine-tuned models. Interestingly, Base+RAG performed inconsistently across frameworks, showing a 5.78% drop in DeepEval but a +3.24% gain in Langchain, reinforcing the variability in retrieval augmentation effects.

Despite these metric-based trends, nonparametric statistical tests revealed no significant differences. The Friedman test produced p values of 0.1218 (DeepEval) and 0.5164 (Langchain), indicating no global performance differences among the four models. Pairwise Wilcoxon signed-rank tests (with a Bonferroni-corrected $\alpha = 0.0083$) similarly yielded nonsignificant results across all comparisons in both evaluation schemes. These findings highlight a key insight: while percentage shifts may suggest directionally meaningful patterns, they were not statistically robust in our current dataset. This suggests the need for larger sample sizes and training the LLM on a larger and more diverse medical dataset without the RAG score. Retrieval augmented generation likely contributes to poorer relevancy outputs across tests. The finetuned LLM without the RAG score performs better than the base model on the above tests and shows consistent directional improvement in both the Langchain and DeepEval tests, although these differences are not statistically significant.

4. Conclusion

- Fine-tuning (FT) is the most consistently beneficial upgrade, especially without a RAG.
- RAG alone provides mixed results. It helps general language slightly but may harm it by likely contributing noise to the output.
- Combining FT + RAG often decreases performance, and focusing on fine-tuning with a larger, high-quality medical dataset will likely improve performance.

5. 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
9. Stevens and Dongfu Jiang and Weiming Ren and Yuxuan Sun and Cong Wei and Botao Yu and Ruibin Yuan and
10. Renliang Sun and Ming Yin and Boyuan Zheng and Zhenzhu Yang and Yibo Liu and Wenhao Huang and Huan
11. Sun and Yu Su and Wenhu Chen. MMMU: A massive multidisciplinary multimodal understanding and
12. Reasoning Benchmark for Expert AGI. https://arxiv.org/abs/2311.16502

Abbreviations

Abbreviation	Definition
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