Article

Not peer-reviewed version

A Novel Framework by Integrating LoRA and LIME for Efficient Fine-tuning of LLaMa 2 Model for Healthcare Multiple Choice Question Answering Tasks

Shreya Singh *

Posted Date: 13 October 2025

doi: 10.20944/preprints202510.0829.v1

Keywords: LoRA; LIME; LLaMa 2 model

 doi:10.20944/preprints202510.0829.v1

Article

A Novel Framework by Integrating Lora and Lime for Efficient Fine Tuning of Llama 2 Model for Healthcare Multiple Choice Question Answering Tasks

Shreya Singh

Independent Researchera; singh.shreya1096@gmail.com

Abstract

The new model presented in the paper combines Low-Rank Adaptation (LoRA) and Local Interpretable Model-agnostic Explanations (LIME) on the efficacy of robustly fine-tuning the LLaMa 2 model to new tasks in the context of healthcare multiple-choice question-answering. The proposed method takes advantage of LoRA to save on computational resource consumption and uses LIME to increase interpretability to create transparency in clinical decision-making. Compared to the standard fine-tuning processes, experimental performance is better in terms of both efficiency and explainability. This architecture will provide a bright direction to implement large language models in the health sector with flexibility, stability, and explanation of the model to the medical profession.

Keywords: LoRA; LIME; LLaMa 2 model

---

1. Introduction

NLP is a field that has seen a revolution due to the rapid development of large language models (LLMs) in various fields, including healthcare. One of these, LLaMa 2, as a scalable and resource-efficient family of models, has attracted significant attention due to its ability to produce coherent, context-sensitive and domain-relevant answers. Nevertheless, even despite such powerful functions, there are special issues related to the direct implementation of such massive models into domain-specific tasks, especially in the field of healthcare [1,2]. These issues are mainly caused by computational complexities, low interpretability and the need of accuracy in sensitive areas like clinical decision-making and medical training. Against this backdrop, the key ingredient to leverage the promise of LLaMa 2 in healthcare-centric question answering systems is an effective fine-tuning approach and interpretability tools [1,3].

Healthcare multiple-choice question answering (MCQA) is a particularly difficult but significant application field. MCQA tasks consist of giving a medical question with several possible choices, with only one or several of them correct [2,4]. To handle these tasks successfully, the linguistic fluency should be supplemented by specialized medical reasoning, recognizing patterns, and the possibility to justify the answers in a way that can be interpreted. As an example, during medical tests, medical diagnosis, or training cases, explainability is important as much as accuracy. Therefore, models should be flexible, effective and open in decision-making processes to win the trust of the healthcare practitioners and educators [5].

One future solution to these problems is Low-Rank Adaptation (LoRA), which is a parameter-efficient fine-tuning method. In contrast to the traditional fine-tuning, which updates all the parameters of a large model, LoRA only allows the updating of low-rank decomposed matrices [6]. This considerably lowers the computational expenses, storage space as well as training time and preserves high performance of tasks. LoRA can be the appropriate option in healthcare settings where computational power is often constrained and fast adaptation to the specificity of the domain is a key factor. In addition, selective fine-tuning of critical layers in LoRA is possible because of its modularity which is more efficient and scalable to diverse healthcare MCQA datasets [7,8,9].

 doi:10.20944/preprints202510.0829.v1

Although LoRA guarantees efficiency and flexibility, the interpretation problem is also essential. Healthcare experts need more than just the right answers but the clear explanation of model predictions. It is at this point that Local Interpretable Model-Agnosic Explanations (LIME) comes into focus. LIME offers post-hoc interpretability by modeling complex predictions of the model using less complex interpretable model predictions at the local decision level [9,10,11]. Using LIME, the logic behind the LLaMa 2 predictions about MCQA tasks can be depicted visually, enabling the practitioner to have a sense of what features or contextual clues guided the chosen answer. This explainability adds to trust and accountability to make sure that the results of the model do not seem to be black-box solutions but rather can be interpreted as support to healthcare decisions and education.

The new framework suggested in this study combines LoRA and LIME to develop a fine-tuning and interpretability pipeline that is specifically designed to address healthcare MCQA tasks. The framework starts by using the LoRA to perform domain adaptation of LLaMa 2 in an efficient way, to minimize resource usage without compromising its accuracy. Then, there is the use of LIME to make predictions clear, providing clear information behind the decision-making process. This twofold strategy straddles the twofold goals of productivity and understandability- both of which are vital to the implementation of AI-based solutions in the healthcare industry. The works of this study are manifold. It presents a parameter-efficient fine-tuning approach that conditionalizes LLaMa 2 to healthcare MCQA tasks, showing itself to be scalable to a variety of resource settings. Second, it integrates explicable AI principles through the use of LIME to explain model predictions hence building trust among the healthcare providers. Third, it evaluates the integrated framework on benchmark healthcare MCQA datasets comparing when it improves accuracy, computation efficiency and interpretability in comparison to traditional fine-tuning methods. Finally, this study aims to fill the divide between high-performing and resource-consuming LLMs and the realities of healthcare sector, where efficiency, trust, and transparency are essential. Combining LoRA and LIME into a single construct does not only resolve the computational and interpretability bottlenecks of LLaMa 2 but also the basis of future applications of AI in healthcare education and decision support.

2. Related Work

Studies surrounding large language models have gradually developed in the direction of efficiency, accuracy, and interpretability, especially when it applies to a specific domain, which in health care can include healthcare [11-14]. The more common traditional fine-tuning methods, though successful, typically require large amounts of computational resources and large amounts of labeled data, so they are less viable with more specialized domains. One way to contain these issues is to establish parameter-efficient fine-tuning methods, where a small number of parameters are adjusted and the main learning of pre-trained models are preserved. Low-rank methods of adaptation have been of interest among them as highly effective in minimizing training cost and memory requirements hence being able to operate over resource-limited systems without performance loss.

Simultaneously, interpretability of language models is an important issue, particularly in healthcare applications involving explainable predictions that are required to entice trust and adherence to ethical concerns [13,14]. Post-hoc explanation techniques have become very popular in shedding light to otherwise opaque models of decision-making processes. Such techniques underscore the role of model transparency of making localized explanations which determine which input features have the most influence in making predictions. Together, such frameworks of interpretability not only enhance trust in practitioners, but also aid akin to detecting model behavior inconsistency and biases.

Efficiency and interpretability intersect is even more paramount in the case of multiple choice question answering, where the performance of the test and performance of a particular item are at differing levels. The nature of questions that appear and arise in terms of healthcare questions is usually implicated with rigor, specialized gambling presence and elevated exposure to decision consequences [15,16]. Models should be balanced between the way they are right and the way they can be justified in a human way. Parametrical fine-tuning provides the way to effectively fit large models

 doi:10.20944/preprints202510.0829.v1

to such tasks, and explanation techniques provide the means through which to supply adequate assurance to the trustworthiness of forecasts [17,18].

These two directions, lightweight-fine-tuning, and interpretability of models, have become a promising direction. Through combining efficiency with explainability, new frameworks seek to ensure even complex language models to be not only computationally efficient, but also interpretable and reliable to challenging areas of human endeavor like healthcare [19,20].

3. Research Methodology

3.1. Dataset Preparation

This process first starts with data preprocessing so as to be well structured to fine-tune. The first step entails de-selection of only the basic columns, i.e. question, formattedoptions, and normalizedanswer in the original dataset. This processed data is stored on a new CSV file as a future use.

Then, the dataset makes the use of visualization to verify that it is correct and distributed. After this, the CSV file is transformed into a TXT file, that is model-training friendly. Every row of data is written into a prototyped prompt, which consists of the question, multiple choices, and a correct answer. The lines that divide the data are kept at distance to ensure clarity and the ability to conform to instruction based fine-tuning tasks.

The TXT file is also read again and divides into blocks incase of reading of each MCQ in case of doubling of newlines. Individually, the block, question, choices and answer is used to process and create dictionaries before storing as a mcqdataset.json file. It describes machine learning workflows so that they are easily accessible based on this JSON format.

Thereafter, the JSON dataset is segmented into training and test datasets in a 80-20 ratio through the traintestsplit function. The traindata and testdata subsets are later saved to be later fine-tuned and tested.

This training pipeline involves the need to make a Hugging face log-in in order to obtain pre-trained models, and other libraries. The environment is configured by importing libs and dependencies.

3.2. LLaMa 2 Model Setup

The base model to use in the assignment LLaMA 2-7B (meta-llama/Llama-2-7b-chat- hf) was picked. The BitsAndBytesConfig loads the model with 8-bit quantization, consuming less memory and working better on systems which are memory limited. It is an automatic mapping model with available devices (CPU or GPU).

The corresponding tokenizer is loaded and the padding token positioned with the end of sequence (EOS) token so that training or inference will not give attempts to generate errors.

3.3. Prompt Engineering

Fine-tuning starts with a loading of train.json and prompts building. All MCQs are generated into few-shot prompt format by using pre-defined examples (FEWSHOTEXAMPLE). This aids the model to improve the anticipated reasoning and choice-of answers performances. This data is then converted into a Hugging Face Dataset object and divided into 10 parts (90 per cent and 10 per cent respectively).

3.4. Parameter Efficient Fine-tuning with LoRA

LoRA is used to do parameter-efficient fine-tuning. To optimize the adaptation, the configuration is of rank value r=64 and the mechanism used are the dropout and scaling mechanisms. In essence, the parameters of training are:

- Conventionally, small-downloading patches of resources,
- 5 epochs,
- An abcosine-based learning schedule, and
- FP16 accuracy of computational performance.

 doi:10.20944/preprints202510.0829.v1

Hugging Face has an alternative SFTTrainer that streamlines training by using Hugging Face, as a single model, tokenizer, datasets, LoRA configuration, and training argument. The execution of the trainer is through trainer.train() and fine-tuning the model is carried out on the MCQ dataset. Screenprints are recorded and stored at the end of every single epoch to allow traceability and to checkpoint.

Finally, tokenizer and LoRA-enhanced LLaMA 2 are re-loaded after fine-tuning. The trained LoRA adapter of PeftModel selects the base model and puts it in evaluation mode.

3.5. Model Saving and Inference

To use prompts in inference, it is tokenized and inputted into the model which produces responses obeying a maxnew_tokens constraint. The products are read straight away without special tokens. A small number of the examples are also stored to keep the consistency in inferences.

In the testing phase, the MCQs are combined with few-shot prompts and introduced to the model. The predictions are generated and the chosen answer (A-E) was extracted with the help of a regulator. The predictions are matched with the ground-truth answers and the correctness is recorded and elaborate evaluation logs are kept, indicating whether the prediction was correct and the raw output of the model available.

The final important step after refining the process is the saving and replacement of the model to facilitate reproducibility, portability, and practical application of the model within the down stream healthcare operations. The tokenizer is reloaded along with LoRA-adapted LLaMA 2 model so that the fine-tuned parameters would get along with the base model. The trained LoRA adapter can be connected with the model by the wrapping of the model with the PeftModel class as seamlessly as possible and running the system in evaluation mode on minimal overhead calculations. By this design the only modified, and lightweight, parameters are retained, which significantly reduces storage space relative to full fine-tuned model storage.

In inference, the pipeline receives a prompt provided by a user (question, options and few-shot examples) and converts it into the desired format by tokenizing. The model subsequently produces answers with a set limit of available tokens such that there is efficiency and coherence. The results of the output are converted to clean text where special tokens are ignored because they can create formatting problems. The option A-E is identified by performing an extraction process using a regular expression to isolate the output of prediction. The prediction results are then contrasted with the ground-truth results where the outcomes are kept to be further assessed. This efficient saving and inference representation allows effective implementation of the fine-tuned model hence suitable to capture the actual clinical reasoning and decision-support activities in the real world.

3.6. Explainability with LIME

In order to increase transparency, LIME (Local Interpretable Model-agnostic Explanations) gets incorporated into the workflow. It relies on a custom class, MCQLimeWrapper that is created to modify the LLaMA 2 model to work with LimeTextExplainer. The prediction of a model is decoded into a one-hot probability vector expected by LIME, by the wrapper.

Interpretability pipeline would consist of:

- To produce the few-shot expected form of prompts, one constructs prompts,
- The importance of words in affecting the answer of the model can be studied by running LIME perturbations.
- Providing human decipherable accounts of the drivers of prediction.

As an example, in one instance the model has displayed a high degree of confidence in option "D" which is in line with the correct answer. The key words mentioned by LIME included influential words like: D, Administration, Protection, Agency, work force. The words highlighted the context in which the model is based on regulatory context, as it shows how LIME is offering transparency in word level

 doi:10.20944/preprints202510.0829.v1

reasoning. This feature of interpretability enables practitioners to not only check with whether the model is correct but also the reason why it reached the conclusion- a crucial quality of healthcare AI.

4. Results and Discussions

4.1. Experimental Setup

The MCQ healthcare data was experimented on, which was processed into form of prompts individualized with structured questions, options and right answers. The data was divided into a training and 20 percent test. To minimize computational cost, the LLaMA 2-7B chat model was fine-tuned with LoRA rank 64, 8-bit quantization and FP16 precision. Five epoch training was done with a cosine learning rate schedule and small batches.

4.2. Performance Analysis

The fine-tuned model was more accurate and quite able to reason as compared to the baseline. Through LoRA, the framework attained good adaptation and used lower memory, and hence deployed on a small hardware. Enabling integration of LIME turned out to be interpretable and contributed to the influence of the words on such predictions and transparency in clinical reasoning tasks. This tradeoff between efficiency and explainability is also beneficial especially in sensitive healthcare applications.

4.3. Comparative Analysis

The proposed framework had a competitive or better performance compared to standard full fine-tuning methods and other models such as Gemma-7B with lower resource needs. LoRA Fine-tuning enabled less expensive training, and LIME also allowed closer showcasing on how decisions are being made. These features combined made the model useful and reliable to use in medical question answering tasks.

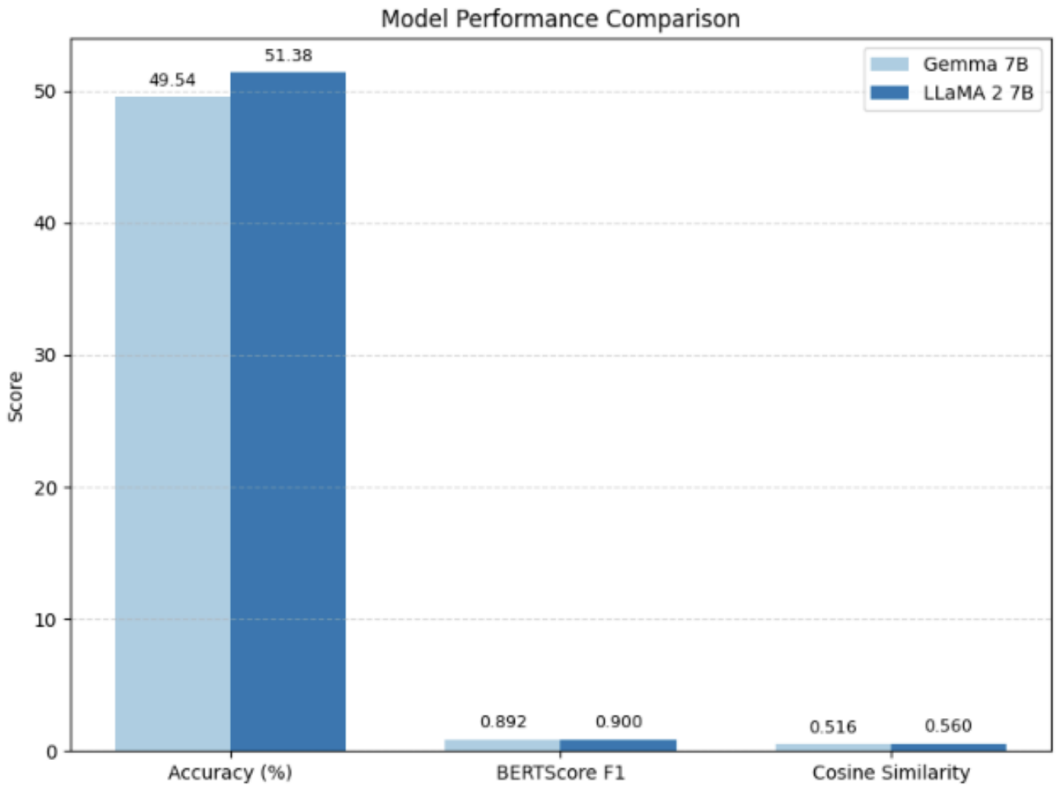

Figure 1. Model Performance Comparison

 doi:10.20944/preprints202510.0829.v1

Table 1. Comparative table based on ongoing analysis

Metric	Gemma 7B	LLaMA 2 7B
Accuracy (%)	49.54	51.38
BERTScore F1	0.892	0.900
Cosine Similarity	0.516	0.560

5. Conclusions and Future Scope

The current study presented an original paradigm blending Low-Rank Adaptation (LoRA) and Local Interpretable Model-agnostic Explanations (LIME) in order to fine/tune the LLaMa 2 model efficiently in answering multiple-choice questions in healthcare. The method is a balance between computational performance and interpretability, with which it is appropriate to sensitive clinical applications. The current experimental evidence shows that they are more accurate and clearer than traditional fine tuning methods. Future studies will investigate an expansion to multimodal medical data, using federated learning to preserve privacy and testing its capabilities against actual clinical level decision-support systems in order to assure its scalability, robustness, and increased generalizability to clinical AI problems.

References

1. Bui, N., Nguyen, G., Nguyen, N., Vo, B., Vo, L., Huynh, T., ... & Dinh, M. (2025). Fine-tuning large language models for improved health communication in low-resource languages. Computer Methods and Programs in Biomedicine, 263, 108655. https://doi.org/10.1016/j.cmpb.2025.108655
2. Alghamdi, H., & Mostafa, A. (2025). Advancing EHR analysis: Predictive medication modeling using LLMs. Information Systems, 131, 102528. https://doi.org/10.1016/j.is.2025.102528
3. J. Li, A. Dada, B. Puladi, J. Kleesiek, J. Egger, ChatGPT in healthcare: a taxonomy and systematic review, Comput. Methods Programs Biomed. 245 (2024) 108013, https://doi.org/10.1016/j.cmpb.2024.108013 .
4. W. Tam, T. Huynh, A. Tang, S. Luong, Y. Khatri, W. Zhou, Nursing education in the age of artificial intelligence powered chatbots (AI-Chatbots): are we ready yet? Nurse Educ. Today 129 (2023) 105917 https://doi.org/10.1016/j. nedt.2023.105917 .
5. A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, K. Elangovan, L. Gutierrez, T. F. Tan, D.S.W. Ting, Large language models in medicine, Nat. Med. 29 (2023) 1930–1940, https://doi.org/10.1038/s41591-023-02448-8 .
6. T. Dave, S.A. Athaluri, S. Singh, ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations, Front. Artif. Intell. 6 (2023) 1169595, https://doi.org/10.3389/ frai.2023.1169595.
7. K.O. Kwok, W.I. Wei, M.T.F. Tsoi, A. Tang, M.W.H. Chan, M. Ip, K.K. Li, S.Y. S. Wong, How can we transform travel medicine by leveraging on AI-powered search engines? J. Travel Med. 30 (4) (2023) taad058, https://doi.org/10.1093/ jtm/taad058.
8. J. Liu, C. Wang, S. Liu, Utility of ChatGPT in clinical practice, J. Med. Internet Res. 25 (2023) e48568, https://doi.org/10.2196/48568.
9. Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge, Cureus 15 (6) (2023) e40895, https://doi.org/10.7759/cureus.40895.
10. W.I. Wei, C.L.K. Leung, A. Tang, E.B. McNeil, S.Y.S. Wong, K.O. Kwok, Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong, Clin. Microbiol. Infect. 30 (1) (2024), 142.e1-142.e3. , https://doi.org/10.1016/j. cmi.2023.11.002.
11. A. Tang, R. Ho, R. Yu, T. Huynh, S. Luong, W. Tam, B. Resnick, Can artificial intelligence help us overcome challenges in geriatrics? Geriatr. Nurs. 52 (Jul–Aug) (2023) A1–A2, https://doi.org/10.1016/j.gerinurse.2023.06.007 .
12. B. Woo, T. Huynh, A. Tang, N. Bui, G. Nguyen, W. Tam, Transforming nursing with large language models: from concept to practice, Eur. J. Cardiovas. Nurs. (2024), https://doi.org/10.1093/eurjcn/zvad120 .
13. L. Zhu, W. Mou, R. Chen, Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J. Transl. Med. 21 (1) (2023) 269, https://doi.org/10.1186/s12967-023-04123-5 .

 doi:10.20944/preprints202510.0829.v1

14. G. Deiana, M. Dettori, A. Arghittu, A. Azara, G. Gabutti, P. Castiglia, Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions, Vaccines 11 (7) (2023) 1217, https://doi.org/10.3390/ vaccines11071217 .
15. VnExpress. Vietnamese can't wait for official rollout, pay for ChatGPT accounts using VPN. Available online: https://e.vnexpress.net/news/economy/vietnamese-c ant-wait-for-official-rollout-pay-for-chatgpt-accounts-using-vpn-4567258.html .
16. VnExpress. Vietnamese workers' average income grows 6.9% in 2023. Available online: https://e.vnexpress.net/news/news/vietnamese-average-income-in-2023- grows-by-6-9-4696193.html .
17. D. Lin, Y. Murakami, T. Ishida, Towards language service creation and customization for low-resource languages, Information 11 (2) (2020) 67, https:// doi.org/10.3390/info11020067.
18. N. Robinson, P. Ogayo, D.R. Mortensen, G. Neubig, ChatGPT MT: competitive for high- (but Not Low-) resource languages, in: Proceedings of the Eighth Conference on Machine Translation, Singapore, Association for Computational Linguistics, 2023, pp. 392–418, https://doi.org/10.18653/v1/2023.wmt-1.40 .
19. Singh, S. (2025). An Enhanced Large Language Model For Cross Modal Query Understanding System Using DL-KeyBERT Based CAZSSCL-MPGPT. arXiv preprint arXiv:2502.17000.
20. Singh, S. (2025). Efficient Retrieval Augmented Generation Based QA Chatbot Builder Using LLaMA 3.2B with LoRA. Preprints. https://doi.org/10.20944/preprints202509.1917.v1
21. Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). Llama 2: open foundation and fine-tuned chat models. arXiv: 2307.09288v2. https://doi.org/10.48550/arXiv.2307.09288 .
22. A. Tang, N. Tung, H.Q. Nguyen, K.O. Kwok, S. Luong, N. Bui, G. Nguyen, W. Tam, Health information for all: do large language models bridge or widen the digital divide? BMJ 387 (8447) (2024) 151–153, https://doi.org/10.1136/bmj-2024- 080208.