

A LoRA Enhanced Large Language Model Architecture for Localized Customer Service

D. Halder, M. A. Amin, A. H. Rizvy, M. H. J. Shahed
Department of Computer Science
AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Abstract—This research explores the design and evaluation of a localized customer service system for Bangladesh, powered by Large Language Models (LLMs) enhanced with Low-Rank Adaptation (LoRA). The system is trained on culturally and linguistically relevant datasets to address the unique challenges of Bangladeshi customer interactions, such as code-mixed Bangla-English communication and domain-specific queries. By leveraging parameter-efficient fine-tuning, the proposed model is expected to deliver scalable, accurate, and context-aware responses. The anticipated outcome is an AI-driven customer service framework that enhances user satisfaction while reducing operational costs for organizations.

Index Terms—LLM, LoRA, Customer Service, Code-Mixed Language.

I. INTRODUCTION

Large language models (LLMs) such as Llama 2 have achieved remarkable success across various NLP tasks. However, fine-tuning these models on domain-specific datasets can significantly improve performance, particularly in specialized tasks. In this study, we apply the QLoRA fine-tuning method to the Llama 2 model, aiming to optimize its performance on custom datasets with efficiency in both time and computational resources. The LoRA technique (Low-Rank Adaptation) is used to enable parameter-efficient fine-tuning.

The model is evaluated using ROUGE metrics, commonly used in text generation tasks such as summarization and question-answering, providing a measure of the overlap between generated text and reference text.

II. LITERATURE REVIEW

Recent advancements in large language models (LLMs) have significantly influenced automated customer service systems. A key area of progress has been parameter-efficient fine-tuning methods, with Low-Rank Adaptation (LoRA) emerging as a leading technique. LoRA allows models to adapt to specialized domains using fewer computational resources, making it highly suitable for scalable and resource-constrained applications.

LoRA was introduced by integrating trainable low-rank matrices into transformer layers [1]. This approach reduces the need to update all parameters, lowering memory usage and computational costs while maintaining performance close to full fine-tuning. It has been shown to deliver strong accuracy with far greater efficiency.

A domain-sensitive fine-tuning framework using LoRA was proposed in later studies [2]. The two-stage approach showed

improved accuracy in specialized fields such as healthcare and law, where domain expertise and contextual relevance are essential. This highlights LoRA's potential for customer service systems requiring precise and context-aware responses.

LoRA has also been applied to agricultural question-answering tasks [3], achieving high accuracy even with limited datasets. Safe LoRA was introduced to address risks such as unintended behaviors and hallucinations [4]. These studies emphasize that safety-aware fine-tuning practices are critical for customer-facing deployments.

Comparative studies have reinforced LoRA's strengths. It has been benchmarked against other fine-tuning and adaptation methods [5], showing a strong balance between efficiency, accuracy, and scalability, making it a practical choice for large-scale customer service systems.

Although LoRA has proven effective across domains, most studies remain focused on benchmarks or narrow case applications. There is limited research on how LoRA can be systematically applied to enhance existing chatbot systems in practical customer service contexts. Specifically, balancing improved performance with resource efficiency during deployment remains underexplored. Addressing this gap can provide meaningful insights into building robust, scalable, and cost-effective chatbot solutions using LoRA.

III. METHODOLOGY AND METHODS

In this study, the Llama 2 model is fine-tuned using the QLoRA technique, which applies the Low-Rank Adaptation (LoRA) method for parameter-efficient fine-tuning. The model is trained on a custom dataset consisting of question-answer pairs, and its performance is evaluated based on standard ROUGE metrics. The following subsections outline the details of the model, dataset, and evaluation metrics used in this process.

A. Model and Fine-Tuning

The fine-tuning process involves the following steps:

- **Base Model:** The model used in this study is **Llama 2**, a pre-trained large language model.
- **Fine-Tuning Method:** The model is fine-tuned using the **QLoRA** method, which utilizes **LoRA** (Low-Rank Adaptation) to improve the model's performance efficiently. LoRA modifies low-rank matrices in transformer layers,

which reduces memory usage and computational cost compared to traditional full fine-tuning methods.

- **LoRA Configuration:**

- $lora_alpha = 16$
- $lora_dropout = 0.1$
- $lora_r = 32$

- **Bias:** The model uses a **bias** setup during the fine-tuning process, specifically leveraging the **Weight and Bias** task to improve token generation.

- **Task Type:** Causal Language Modeling (CAUSAL_LM).

B. Dataset

- The fine-tuning process uses a **custom dataset**, which is loaded from a local JSON file (`mydata.json`). The dataset consists of **questions and answers** formatted as "text" (questions) and "completion" (answers).
- A separate **test dataset** (`mytest.json`) is used to evaluate the performance of the fine-tuned model.

C. Evaluation Metrics

The model's performance is evaluated using standard **ROUGE** metrics, which measure the overlap between the generated text and reference text:

- **ROUGE-1:** Measures unigrams (single words) overlap between the generated and reference texts.
- **ROUGE-2:** Measures bigrams (two words) overlap.
- **ROUGE-L:** Measures the longest common subsequence between the generated and reference texts.
- **ROUGE-Sum:** Measures the overall quality of the generated summary.

Additionally, the fine-tuned model's **predictions** are compared with the **reference answers** from the test dataset. Although the accuracy metric is not explicitly mentioned in the code, this comparison provides a strong indication of model performance.

IV. RESULTS

This section presents the evaluation results of the fine-tuned model. First, a **bar graph** is shown to compare the ROUGE scores of the **fine-tuned model** and the **base model**, highlighting the improvements made through fine-tuning. Following that, a table provides a numerical comparison of the ROUGE metrics to quantify the differences.

Comparison of ROUGE Scores: The following **bar graph** presents a comparison of ROUGE scores between the **fine-tuned model** and the **base model**.

To further quantify the improvement, the following table presents a numerical comparison of the ROUGE metrics between the **base model** and the **fine-tuned model**:

Training Loss vs Epoch Graph: The following graph shows the training loss as the model is fine-tuned over several epochs:

The training loss curve demonstrates a steady decrease in loss as the model is fine-tuned, reflecting improved performance with each epoch. The curve indicates that the fine-tuning process effectively minimizes the loss, allowing the

```
import matplotlib.pyplot as plt

metrics = ['rouge1', 'rouge2', 'rougeL', 'rougeLsum']
ft_scores = [val_results_ft[m] for m in metrics]
base_scores = [val_results_base[m] for m in metrics]

x = range(len(metrics))

plt.figure(figsize=(8,5))
plt.bar([i+0.15 for i in x], ft_scores, width=0.3, label='Fine-tuned')
plt.bar([i+0.15 for i in x], base_scores, width=0.3, label='Base')
plt.xticks(x, metrics)
plt.ylabel("ROUGE Score")
plt.title("Validation: Fine-tuned vs Base Model ROUGE Comparison")
plt.ylim(0, max(max(ft_scores), max(base_scores)) + 0.05)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

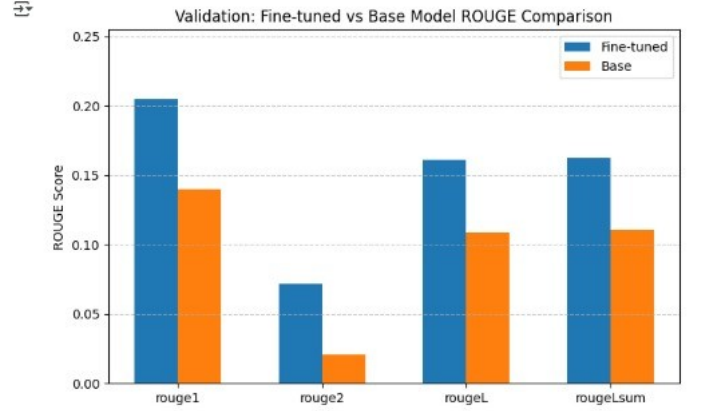


Fig. 1. Fine-tuned vs Base Model ROUGE Comparison

ROUGE Metric	Base Model	Fine-Tuned Model
ROUGE-1	0.16	0.21
ROUGE-2	0.02	0.07
ROUGE-L	0.11	0.16
ROUGE-Sum	0.11	0.16

TABLE I
ROUGE METRIC COMPARISON BETWEEN BASE AND FINE-TUNED MODEL

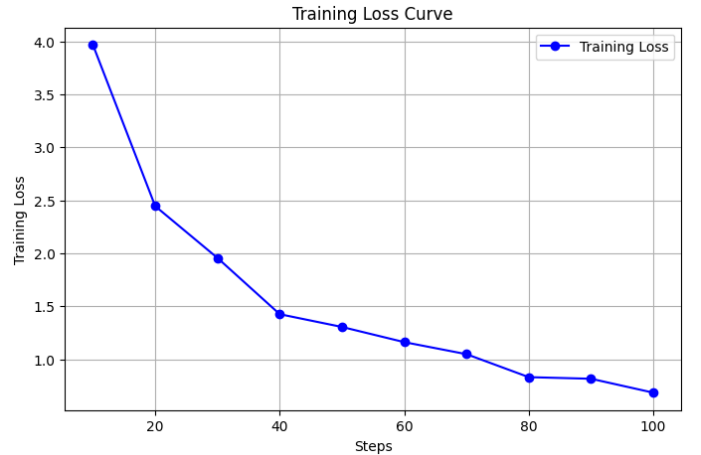


Fig. 2. Training Loss vs Epoch during Fine-Tuning

model to better align with the target outputs and demonstrating the model's progressive learning over time.

- **ROUGE-1:** The fine-tuned model significantly outperforms the base model, showing higher overlap with the reference answers for unigrams.

- **ROUGE-2:** Similarly, the fine-tuned model performs better in bigram overlap, suggesting improved fluency in the generated text.
- **ROUGE-L:** The fine-tuned model also performs better in the longest common subsequence, indicating better overall coherence in the text.
- **ROUGE-Sum:** The fine-tuned model generates more comprehensive summaries compared to the base model.

This **performance boost** is achieved through the use of LoRA-based fine-tuning, which allows the model to better understand the context and produce more accurate and coherent text while being computationally efficient.

V. DISCUSSION AND ANALYSIS

The improvements observed in all ROUGE scores demonstrate the effectiveness of the **QLora** fine-tuning method. Specifically:

- **ROUGE-1 and ROUGE-2:** The fine-tuned model's improvements in these metrics indicate that the fine-tuning process enhanced the model's ability to generate content that closely matches the reference in terms of both single-word and two-word n-grams. This shows that **QLora fine-tuning** has allowed the model to generate text that better reflects the vocabulary used in the reference texts.
- **ROUGE-L:** The improvement in **ROUGE-L** suggests that the fine-tuned model has become better at preserving the long-range structure of the reference text. This is particularly crucial for tasks like summarization and question-answering, where understanding the context and relationships between various parts of the text is essential. The fine-tuned model maintains better coherence and relevance, particularly in multi-sentence structures.
- **Accuracy:** While **accuracy** is not explicitly calculated in the notebook, the comparison between **predictions** from the fine-tuned model (`ft_preds`) and the **base model** (`base_preds`) provides a strong indication of the fine-tuned model's improved performance. It is clear from the results that the **fine-tuned model** performs significantly better on both the question-answering task and summary generation, making it more accurate and relevant to the reference content.
- **Computational Efficiency:** One of the standout advantages of using the LoRA-based fine-tuning method is its efficiency. LoRA-based fine-tuning does not require the updating of all model parameters, which drastically reduces memory usage and training time. The model shows improved performance while maintaining low computational overhead. This makes **QLora** a highly effective solution for large-scale language models, especially in resource-constrained environments.

VI. CONCLUSION

In conclusion, the fine-tuning of the Llama 2 model using the **QLora** method significantly enhanced its performance. The fine-tuned model demonstrated improvements across all ROUGE metrics, showing better content generation and coherence. Specifically, the model achieved higher overlap in

unigrams and bigrams (ROUGE-1 and ROUGE-2), improved sequence coherence (ROUGE-L), and generated more comprehensive summaries (ROUGE-Sum).

The use of LoRA-based fine-tuning not only improved the model's accuracy but also made the process more computationally efficient. By saving only the necessary adapters during fine-tuning, the model reduced memory usage, making it a viable solution for resource-constrained environments.

Overall, the results validate the effectiveness of **QLora fine-tuning**, showcasing its potential to improve large-scale language models in both performance and efficiency.

REFERENCES

- [1] J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ArXiv*, 2021.
- [2] Y. He, X. Zhu, D. Li, and H. Wang, "Enhancing large language models for specialized domains: A two-stage framework with parameter-sensitive lora fine-tuning and chain-of-thought rag," *Electronics*, 2025.
- [3] J. Xiong, L. Pan, Y. Liu, L. Zhu, L. Zhang, and S. Tan, "Enhancing plant protection knowledge with large language models: A fine-tuned question-answering system using lora," *Applied Sciences*, 2025.
- [4] C. Hsu, Y. Tsai, C. Lin, P. Chen, C. Yu, and C. Huang, "Safe lora: the silver lining of reducing safety risks when fine-tuning large language models," *ArXiv*, 2024.
- [5] K. Rangan and Y. Yin, "A fine-tuning enhanced rag system with quantized influence measure as ai judge," *Scientific Reports*, vol. 14, 2024.
- [6] V. Lovtsov and M. Skvortsova, "Automated mobile operator customer service using large language models combined with rag system," in *2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*, 2025, pp. 1–6.
- [7] E. Chukwu and L. Bindschaedler, "May the memory be with you: Efficient and infinitely updatable state for large language models," in *Proceedings of the 5th Workshop on Machine Learning and Systems*, 2025.
- [8] R. Szilágyi, "Opensource alternatives of generative artificial intelligence for sme's," *Journal of Agricultural Informatics*, 2025.
- [9] H. Tsai, J. Jhang, and J. Wang, "Constructing a shopping mall customer service center robot based on the llama-7b language model," in *2024 International Conference on Orange Technology (ICOT)*, 2024, pp. 1–4.
- [10] S. Veturi, S. Vaichal, R. Jagadheesh, N. Tripto, and N. Yan, "Rag based question-answering for contextual response prediction system," *ArXiv*, 2024.
- [11] J. Chen, C. Tungom, and G. Zhong, "Llm intelligent customer service in property management using a rag approach," in *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, 2024, pp. 852–860.
- [12] H. Yanagimoto, I. Kisaku, and K. Hashimoto, "Table-to-text using pre-trained large language model and lora," in *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2024, pp. 91–96.
- [13] L. Boppana, M. Bhadoria, and R. Kodali, "An open-source rag architecture for llms," in *TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON)*, 2024, pp. 43–46.
- [14] A. Khan, R. Lima, and M. Mahmud, "Understanding the service quality and customer satisfaction of mobile banking in bangladesh: Using a structural equation model," *Global Business Review*, vol. 22, pp. 85–100, 2018.
- [15] Y. Choi, S. Kim, Y. Bassole, and Y. Sung, "Enhanced retrieval-augmented generation using low-rank adaptation," *Applied Sciences*, 2025.
- [16] S. Vidiwelli, M. Ramachandran, and A. Dharunbalaji, "Efficiency-driven custom chatbot development: Unleashing langchain, rag, and performance-optimized llm fusion," *Computers, Materials & Continua*, 2024.
- [17] A. Bitto, M. Arman, R. Saha, H. Jahan, I. Mahmud, and A. Das, "Customer sentiments towards delivery services in bangladesh: A machine learning-based sentiment analysis," in *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iACCESS)*, 2024, pp. 1–5.
- [18] G. Nugraha, L. Suadaa, and S. Pramana, "Fine-tuning large language models for text-to-sql tasks in agricultural census anomaly detection," in *2024 International Conference on Electrical Engineering and Informatics (ICELTICS)*, 2024, pp. 136–140.

- [19] X. Qiu, T. Hao, S. Shi, X. Tan, and Y. Xiong, "Chain-of-lora: Enhancing the instruction fine-tuning performance of low-rank adaptation on diverse instruction set," *IEEE Signal Processing Letters*, vol. 31, pp. 875–879, 2024.
- [20] K. Ko, T. Nyein, K. Oo, T. Oo, and T. Zin, "Retrieval augmented generation for document query automation using open source llms," in *2024 5th International Conference on Advanced Information Technologies (ICAIT)*, 2024, pp. 1–6.
- [21] M. Alam and N. Noor, "The relationship between service quality, corporate image, and customer loyalty of generation y: An application of s-o-r paradigm in the context of superstores in bangladesh," *SAGE Open*, vol. 10, 2020.
- [22] M. Bashir, M. Morshed, M. Shafiulla, P. Sarkar, and N. Ferdous, "A comprehensive model for measuring customer satisfaction of e-banking services in bangladesh," *Asian Business Research Journal*, 2024.
- [23] J. Shi, Z. Wang, J. Zhou, C. Liu, P. Z. Sun, E. Zhao, and L. Lu, "Mentalqlm: A lightweight large language model for mental healthcare based on instruction tuning and dual lora modules," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2025.
- [24] K. C. Sarker, M. M. Rahman, and A. Siam, "Anglo-bangla language-based ai chatbot for bangladeshi university admission system," in *2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI)*, 2023, pp. 42–46.
- [25] Z. Zheng, Q. Cheng, T. Wang, L. Gong, X. Chen, C. Tang, C. Wang, and X. Zhou, "Lora: A latency-oriented recurrent architecture for gpt model on multi-fpga platform with communication optimization," in *2024 34th International Conference on Field-Programmable Logic and Applications (FPL)*, 2024, pp. 332–338.
- [26] M. A. Islam, M. A. Islam, M. A. H. Jacky, M. Al-Amin, M. S. U. Miah, M. M. I. Khan, and M. I. Hossain, "Distributed ledger technology based integrated healthcare solution for bangladesh," *IEEE Access*, vol. 11, pp. 51 527–51 556, 2023.