2024 International Conference on Orange Technology (ICOT) | 979-8-3315-2360-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOT64290.2024.10936962

Constructing a Shopping Mall Customer Service Center Robot Based on the LLAMA-7B Language Model

Hsin-Chun Tsai*, Jhe-Wei Jhang, and Jhing-Fa Wang
Department of Electrical Engineering, National Cheng-Kung University, Tainan, Taiwan
*Email: tsaihcmail@gmail.com

***Abstract*—This study develops a customer service robot powered by a large language model, deployed at Tainan Spinning Mall. The system runs on the Temi robot and leverages Llama3 fine-tuned using PEFT methods like LoRA to optimize performance on an NVIDIA L4 GPU with limited memory. Retrieval-Augmented Generation (RAG) enhances response accuracy, while training data includes mall-specific datasets, such as store directories, product listings, and promotions, alongside general dialogue data. After model adaptation, it achieves 95% accuracy in answering mall-related inquiries, responding within five seconds, ensuring efficient and accurate assistance in fast-paced retail environments.**

Keywords—Large Language Model (LLM), fine-tuning, customer-service chatbot

I. INTRODUCTION

In shopping malls, the quality of customer service responses is crucial for improving customer satisfaction. However, traditional human customer service faces limitations, such as restricted working hours and incomplete knowledge bases. To address these issues, this study proposes a customer service robot based on a large language model [1], capable of understanding and answering a wide range of customer queries. The Llama3 model was chosen for training due to its highly efficient architecture and superior stability. The research objectives include building a mall-related knowledge base, enhancing system usability through voice interaction interface, and improving response accuracy. Experimental results indicate that the customer service robot performs exceptionally well in answering mall-related questions, proving the potential of large language models in real-world applications.

II. RELATED WORK

In 2018, OpenAI released the GPT-2 model based on the Transformer architecture [2]. The GPT-2 model is capable of performing a variety of NLP tasks, including language translation and dialogue systems. Subsequently, OpenAI and other research institutions continued to advance this technology, launching GPT-3 [3], which further enhanced the capabilities of large language models.

In our system, we also chose a Transformer-based model as the foundation. The advantages of the Transformer include its high parallelism and self-attention mechanism, which enable the model to handle long texts and diverse language tasks with robust capabilities. Additionally, we will explore research on LLM chatbots based on RAG (Retrieval-Augmented Generation) technology [4]. This study used Llama2-7b [5] as the base model, employing PDF files as input data and achieving over 95% accuracy through RAG technology.

In this chapter, we will examine various research methods used in other large language model (LLM) systems, such as Adapter Tuning [6], P-Tuning v2 [7], and attempt to compare different training approaches to optimize our system in the most suitable manner. We will analyze the advantages and disadvantages of these methods and explore how to leverage these research findings to further enhance the performance of the Llama3 model.

III. METHODS

This block diagram shown in Fig.1 mainly includes model training and Retrieval-Augmented Generation (RAG). During the training phase, we enhance the model's summarization capability to achieve better performance in the RAG process. In the RAG phase, the model retrieves text from our TSmall dataset and optimizes it to obtain the best results. We will introduce the following sections: 1. Summarization dataset, 2. TSmall dataset, 3. Model fine-tuning, 4. Retrieval-Augmented Generation, 5. Voice Interaction, and 6. Performance evaluation.

A. Summarization dataset

The data we used in this work is CLTS-dataset, which is a dataset containing Chinese long text summarization examples. We select about 10,000 summary examples for training. The advantage of this dataset is that it consists of Chinese long texts, which is more suitable for our application field than other non-Chinese short-text summarizations.

B. TSmall dataset

The TSmall dataset contains various types of information about the shopping mall, including store details, facilities, events, and policies. To enhance the accuracy of responses, the system integrates Large Language Models (LLMs) with RAG (Retrieval-Augmented Generation) technology, improving the system's response accuracy to shopping mall-related queries. Table I shows the details of TSmall dataset.

This work was supported in part by the National Science and Technology Council of Taiwan, R.O.C., under Grant MOST 113-2221-E-006 -212 -MY2

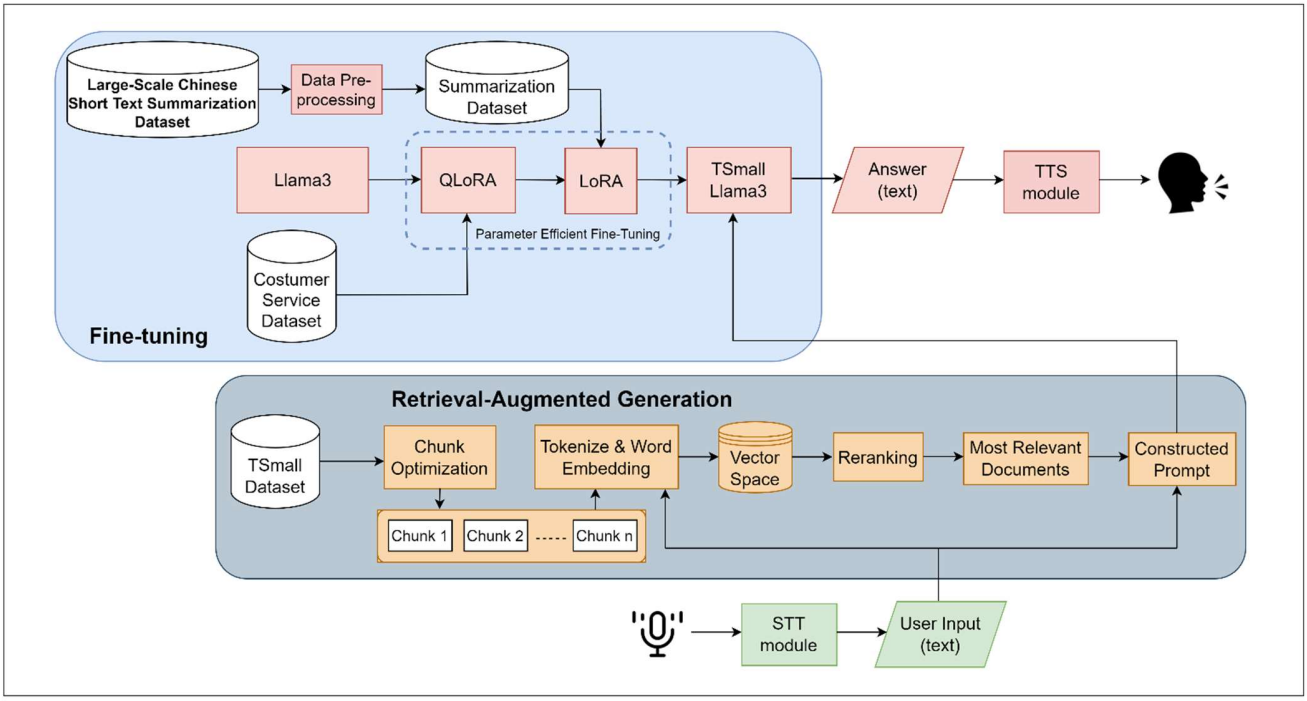

Fig. 1. Overall block diagram

TABLE I
TSMALL DATASET

	Store information	Mall facilities	Activities information	Mall policy
words count	82520	12368	13214	2010

C. Model fine-tuning

To address the issue of large pre-trained models requiring extensive computational resources, we use PEFT technology to reduce the number of fine-tuning parameters and computational complexity, thereby enhancing model performance on new tasks. LoRA [8], as a solution, replaces full-parameter fine-tuning with the insertion of low-rank matrices, reducing computational costs while simultaneously improving the model's capabilities. The actual operation of LoRA is shown in Figure.2, which is provied from [8]

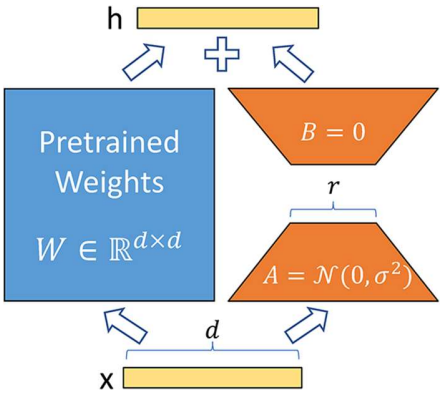

Fig. 2. Operation of LoRA

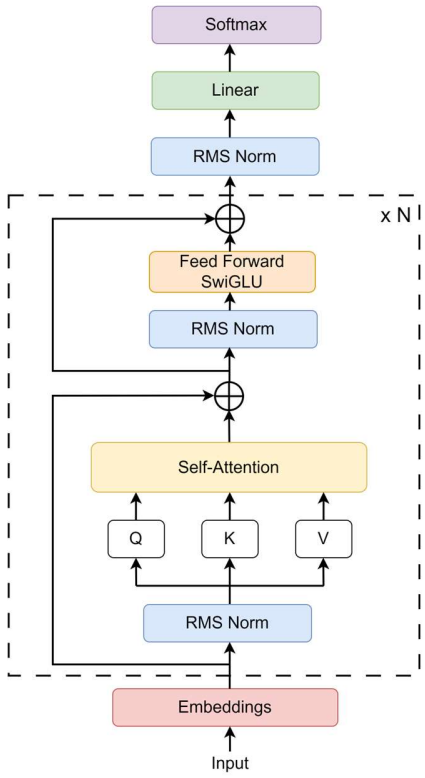

Fig. 3. The structure of Llama3.

D. Llama3

Llama3 is designed based on an advanced Transformer architecture, which has been widely used in the field of natural language processing. Additionally, we have incorporated various data pre-processing techniques and optimization methods to further enhance the model's performance and

stability. We have adjusted the parameters of the Q, K, and V matrices in Llama3 to improve its performance on specific tasks. Fig. 3 shows the architecture of Llama3 model.

E. Retrieval-Augmented Generation

The system utilizes Retrieval Augmented Generation (RAG) [4] technology for information retrieval, quickly finding relevant texts from the knowledge base to assist the Large Language Model (LLM) in generating responses. RAG can flexibly retrieve knowledge to address various queries in mall customer service. Additionally, the system incorporates re-ranking [9] techniques to further improve the accuracy and relevance of the responses.

F. Voice Interaction

STT (Speech-to-Text) uses the Webkit Speech Recognition tool to perform speech-to-text conversion on the frontend, addressing the latency issues caused by backend processing and improving system responsiveness. TTS (Text-to-Speech) employs the Speech Synthesis frontend library to enable real-time text-to-speech conversion.

IV. EXPERIMENTS

NVIDIA L4 GPU with 24GB of memory. The memory size of the L4 GPU is well-suited for the dataset used in this study, enabling the model to be trained without exceeding hardware resource limitations.

Additionally, in this study, the Temi robot is used as the deployment platform, aiming to integrate the system with the service robots commonly found in shopping malls. By applying our system to a service robot, we seek to enhance its practicality and usability in real-world service scenarios.

Fig. 4. Temi robot

B. Experiment of Retrieval Efficiency

In this section, we will introduce the metrics we use to evaluate retrieval efficiency. We employ Hit Rate (1) and MRR(2) (Mean Reciprocal Rank) for this assessment.

Table. II is the experiment result before we added re-ranking. Table. III is the experiment result after we added re-ranking. From the tables of both metrics, it can be observed that when the number of retrievals is 6, the Hit Rate increases from

TABLE II
HIT RATES FOR RETRIEVERS WITHOUT RE-RANKING

k	Hit Rate	MRR
1	0.7490	0.6596
2	0.7679	0.6711
3	0.7810	0.6914
4	0.8017	0.7098
5	0.8272	0.7113
6	0.8384	0.7326

TABLE III
HIT RATES FOR RETRIEVERS WITH RE-RANKING

k	Hit Rate	MRR
1	0.7892	0.7214
2	0.8090	0.7442
3	0.8241	0.7693
4	0.8415	0.7905
5	0.8755	0.8114
6	0.8834	0.8207

0.8384 to 0.9134. This indicates a significant improvement in the retrieval efficiency of the system, which helps reduce the amount of irrelevant information received by the model.0

C. Experiment of Response Accuracy

In the previous chapter, we conducted experiments on the retrieval methods used in this system to identify the most suitable approach for the application scenario discussed in this paper. Next, the model combines the retrieved texts with the query as a new input to generate responses. Therefore, we will now evaluate the similarity between the model-generated responses and the retrieved text segments to test whether the model can provide accurate responses based on the retrieved information.

This system uses BERTScore as the accuracy metric for model responses. BERTScore is a widely used evaluation metric for text comparison tasks, assessing the similarity between generated text and reference text. The evaluation metrics for this method are as follows:

TABLE IV
SCORES FOR DIFFERENT INFORMATION CATEGORIES WITH AND WITHOUT RE-RANKER.

	Store Information	Mall Facilities	Activities Information	Mall Policy
	With Reranker			
Recall	0.82	0.73	0.80	0.74
Precision	0.59	0.71	0.78	0.70
F1 Score	0.69	0.72	0.79	0.72
	Without Reranker			
Recall	0.69	0.67	0.67	0.65
Precision	0.63	0.72	0.71	0.67
F1 Score	0.66	0.69	0.69	0.66

$$\text{Recall} = \frac{\text{Ture Positives}}{\text{True Positives + False Negatives}} \quad (3)$$

$$\text{Precision} = \frac{\text{Ture Positives}}{\text{True Positives + False Positives}} \quad (4)$$

$$\text{F1}_{\text{score}} = \frac{2\text{precision}\cdot\text{recall}}{\text{precision+ recall}} \quad (5)$$

where TP is true positive, FP is false positive, and FN is false negative. Table IV-C shows the evaluation results.

V. CONCLUSION

We use Llama3 as the base model and fine-tune it using the PEFT method to improve training efficiency and reduce the risk of overfitting. Next, we enhance system accuracy by utilizing Retrieval-Augmented Generation (RAG) technology and the TSmall dataset established in this paper. During the RAG process, we combine Re-ranking techniques to further filter retrieved texts, thereby providing more accurate and relevant answers. Finally, we developed and implemented a low-latency voice dialogue interface. This was achieved through the integration of front-end Text-to-Speech (TTS) and Speech-to-Text (STT) technologies, making the voice interaction smoothly and efficiently with minimal delays.

REFERENCES

[1] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Galle´ *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," 2023.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal,A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[7] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," arXiv preprint arXiv:2110.07602, 2021.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[9] C. Pei, Y. Zhang, Y. Zhang, F. Sun, X. Lin, H. Sun, J. Wu, P. Jiang,J. Ge, W. Ou et al., "Personalized re-ranking for recommendation," in Proceedings of the 13th ACM conference on recommender systems, 2019, pp. 3–11.