ABSTRACT

Customer service on campus often faces challenges in providing fast and accurate responses, especially with conventional methods that require long wait times and cannot handle more complex queries. The use of Large Language Models (LLM)-based chatbots offers a solution to improve service efficiency, however, its application in resource-constrained environments faces constraints in memory and computational efficiency. This research develops an LLM-based customer service chatbot optimised using the Quantized Low-Rank Adaptation (QLoRA) technique to address memory efficiency issues in a campus environment. The Mistral-7B model was fine-tuned using the QLoRA technique, which reduces memory usage through 4-bit quantisation, thus enabling the use of large models in resource-constrained environments. Experimental results show that the model fine-tuned using QLoRA is able to provide accurate and relevant responses with BLEU values of 0.6992 and ROUGE of 0.8659, and is efficient in resource usage. This research successfully demonstrated the application of LLM-based chatbot with QLoRA can improve the quality of customer service on campus by optimally utilising limited resources.

Keywords: *chatbot*, *large language models*, QLoRA, *customer service*, *fine-tuning*.