ABSTRAK

Layanan pelanggan di kampus sering kali mengalami tantangan dalam menyediakan respons yang cepat dan akurat, terutama dengan metode konvensional yang memerlukan waktu tunggu panjang dan tidak dapat menangani pertanyaan yang lebih kompleks. Penggunaan *chatbot* berbasis *Large Language Models* (LLM) menawarkan solusi untuk meningkatkan efisiensi layanan, namun penerapannya pada lingkungan sumber daya terbatas menghadapi kendala dalam efisiensi memori dan komputasi. Penelitian ini mengembangkan *chatbot* layanan pelanggan berbasis LLM yang dioptimalkan menggunakan teknik *Quantized Low-Rank Adaptation* (QLoRA) untuk mengatasi masalah efisiensi memori dalam lingkungan kampus. Model Mistral-7B di-*fine-tune* dengan teknik QLoRA, yang mengurangi penggunaan memori melalui kuantisasi 4-bit, sehingga memungkinkan penggunaan model besar pada lingkungan sumber daya terbatas. Hasil eksperimen menunjukkan bahwa model yang di-*fine-tune* menggunakan QLoRA mampu memberikan respons akurat dan relevan dengan nilai BLEU 0.6992 dan ROUGE 0.8659, serta efisien dalam penggunaan sumber daya. Penelitian ini berhasil menunjukkan penerapan *chatbot* berbasis LLM dengan QLoRA dapat meningkatkan kualitas layanan pelanggan di kampus dengan memanfaatkan sumber daya terbatas secara optimal.

Kata Kunci: *chatbot*, *large language models*, QLoRA, layanan pelanggan, *fine-tuning*.