INTISARI

Pengembangan *chatbot* berbasis *Large Language Model* (LLM) telah mengubah interaksi digital dengan komputer menjadi semakin mirip dengan interaksi manusia, namun efisiensi sumber daya komputasi dan akurasi respons masih menjadi tantangan. Penelitian ini berfokus pada analisis perbandingan beban komputasi dan akurasi *chatbot* LLM dengan implementasi *Retrieval Augmented Generation* (RAG) dan NON-RAG. Sistem *chatbot* dibangun menggunakan Ollama sebagai LLM lokal, Qdrant sebagai basis data vektor, PostgreSQL untuk *chat log*, dan n8n sebagai pusat *workflow*, seluruhnya di-deploy dalam Docker. Metodologi penelitian meliputi pengujian beban komputasi (CPU, RAM, waktu eksekusi) pada tiga skenario: *baseline* (tanpa *chatbot*), *chatbot* NON-RAG, dan *chatbot* RAG. Analisis juga mencakup perbandingan penggunaan basis komputasi CPU dan GPU untuk skenario RAG. Selain itu, akurasi jawaban *chatbot* dievaluasi berdasarkan kesesuaian dengan data eksternal. Hasil penelitian menunjukkan perbedaan signifikan dalam beban komputasi dan waktu eksekusi antara skenario NON-RAG dan RAG, serta dampak penggunaan CPU dibanding GPU pada kinerja sistem. Evaluasi akurasi memberikan wawasan tentang peningkatan kualitas respons dengan implementasi RAG. Hasil penelitian ini memberikan data empiris yang penting bagi pengembang dalam perancangan arsitektur dan pemilihan hardware, serta menjadi referensi bagi penelitian selanjutnya di bidang *chatbot* berbasis LLM.

Kata kunci: *Chatbot, Large Language Model (LLM), Retrieval Augmented Generation (RAG)*, Beban Komputasi, Akurasi, n8n.

ABSTRACT

The development of Large Language Model (LLM)-based chatbots has transformed digital interactions with computers, making them increasingly human-like; however, computational resource efficiency and response accuracy remain challenges. This research focuses on analyzing the comparative computational load and accuracy of LLM chatbots with Retrieval Augmented Generation (RAG) and NON-RAG implementations. The chatbot system is built using Ollama as the local LLM, Qdrant as the vector database, PostgreSQL for chat logs, and n8n as the workflow hub, all deployed within Docker. The research methodology includes testing computational load (CPU, RAM, execution time) across three scenarios: baseline (no chatbot), NON-RAG chatbot, and RAG chatbot. The analysis also compares CPU versus GPU computational base usage for the RAG scenario. Additionally, chatbot answer accuracy is evaluated based on its alignment with external data. The results show significant differences in computational load and execution time between the NON-RAG and RAG scenarios, as well as the impact of CPU compared to GPU usage on system performance. The accuracy evaluation provides insights into response quality improvements with RAG implementation. These findings provide crucial empirical data for developers in architecture design and hardware selection, and serve as a reference for future research in LLM-based chatbots.

Keywords: Chatbot, Large Language Model (LLM), Retrieval Augmented Generation (RAG), Computational Load, Accuracy, n8n.