

INTISARI

INTEGRASI *LARGE LANGUAGE MODELS* DALAM RANCANG BANGUN *CHATBOT* BERBASIS *RETRIEVAL-AUGMENTED* *GENERATION* DAN *CLOUD COMPUTING* UNTUK EFISIENSI AKSES INFORMASI PADA WEBSITE PENDUKUNG AKADEMIK

Yana Dayinta Nesthi

21/478358/SV/19272

Chatbot konvensional kerap kali berhadapan dengan tantangan dalam menghasilkan respons yang sesuai konteks dan pengelolaan beban kerja yang berat. Proyek ini mengembangkan *chatbot* pendukung akademik berbasis RAG dan LLMs bernama "Trichie" yang diimplementasikan pada infrastruktur *cloud computing* dan diotomatisasikan menggunakan n8n, guna meningkatkan efisiensi akses informasi di *website* resmi program studi TRI SV UGM.

Dengan tujuan utama untuk mengoptimalkan kinerja Trichie, merancang *workflow* yang efisien, dan mengevaluasi kepuasan pengguna, sistem *back-end* Trichie mampu melakukan proses *crawling* secara otomatis dan juga memproses data berupa teks, gambar, dan PDF. Data ini kemudian disusun menjadi sebuah *knowledge base* di Pinecone menggunakan *embedding* OpenAI. Dengan mengintegrasikan *back-end chatbot* yang ada di n8n dengan Chatwoot sebagai antarmuka penggunaannya, memungkinkan Trichie untuk memberikan respons secara *real-time*. Hasil dari rancangan sistem Trichie berhasil menunjukkan skor *Customer Satisfaction Score* (CSAT) sebesar 4,59/5 dengan waktu respons rata-rata 12,727 detik.

Kata kunci: *Chatbot*, *Retrieval-Augmented Generation*, *Large Language Models*, *Cloud Computing*, n8n, platform otomatisasi alur kerja, Pinecone, Supabase, *Customer Satisfaction Score* (CSAT).

ABSTRACT

INTEGRATION OF LARGE LANGUAGE MODELS IN THE FRAMEWORK OF A RETRIEVAL-AUGMENTED GENERATION AND CLOUD COMPUTING-BASED CHATBOT FOR EFFICIENT INFORMATION ACCESS ON AN ACADEMIC SUPPORT WEBSITE

Yana Dayinta Nesthi

21/478358/SV/19272

Conventional chatbots frequently encounter challenges in generating contextually relevant responses and managing heavy workloads. This project addresses these issues by developing an academic support chatbot named "Trichie," which leverages Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). Trichie is implemented on a cloud computing infrastructure and automated using n8n, aiming to enhance the efficiency of information access on the official website of the TRI SV UGM study program.

The primary objectives are to optimize Trichie's performance, design an efficient workflow, and evaluate user satisfaction. Trichie's back-end system is capable of automatic crawling and processing various data types, including text, images, and PDFs. This data is then organized into a knowledge base in Pinecone using OpenAI embeddings. By integrating the n8n-based chatbot back-end with Chatwoot as its user interface, Trichie can provide real-time responses. The system design for Trichie successfully achieved a Customer Satisfaction Score (CSAT) of 4.59/5, with an average response time of 12.727 seconds.

Keywords: Chatbot, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Cloud Computing, n8n, workflow automation platform Pinecone, Supabase, Customer Satisfaction Score (CSAT)