Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur
Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

IMPLEMENTASI DAN ANALISIS KINERJA LAYANAN *CHATBOT* BERBASIS *LARGE LANGUAGE MODEL* DENGAN ARSITEKTUR *RETRIEVAL AUGMENTED GENERATION* (STUDI KASUS LAYANAN UGM ANJEM)

Anugrah Dwiki Ardana

22/494995/SV/20942

UGM Anjem adalah layanan antar jemput mahasiswa Universitas Gadjah Mada yang selama ini mengandalkan WhatsApp sebagai satu-satunya saluran informasi, sehingga pertanyaan berulang harus dijawab manual oleh admin. Proyek akhir ini mengimplementasikan *chatbot* berbasis RAG menggunakan FastAPI, GPT-4o mini, ChromaDB, dan Supabase yang di-*deploy* pada VPS kampus UGM, serta menganalisis kinerja layanannya menggunakan Prometheus dan Grafana berdasarkan kerangka *Four Golden Signals*. Evaluasi kualitas jawaban dilakukan melalui fitur *rating* (5 responden mahasiswa UGM) dan RAGAS (30 pertanyaan uji), sedangkan pengujian kinerja menggunakan Locust dengan empat skenario: 10 pengguna, 25 pengguna, 50 pengguna, dan 100 pengguna. Hasil menunjukkan *helpful ratio* 92% dengan rata-rata skor 3,58/4,00. RAGAS menghasilkan *faithfulness* 0,748, *answer relevancy* 0,693, dan *answer correctness* 0,639, di mana *faithfulness* yang tinggi mengindikasikan RAG berhasil menekan *hallucination* secara efektif. Dari sisi kinerja, sistem andal pada skenario 10 pengguna dan 25 pengguna dengan *error rate* 0% dan P50 stabil di 3,10 detik. Penurunan kinerja terdeteksi pada skenario 50 pengguna (*error rate* 5,22%) dan semakin parah pada skenario 100 pengguna (*error rate* 28,51%), dengan *bottleneck* utama pada tahap *Generation* oleh OpenAI API, sementara saturasi infrastruktur lokal tetap di bawah 3% CPU dan 10% memori di seluruh skenario.

Kata kunci: *chatbot*, RAG, GPT-4o mini, *service performance analysis*, RAGAS, *load testing.*

UNIVERSITAS
GADJAH MADA
**Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur
Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)**
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

IMPLEMENTATION AND PERFORMANCE ANALYSIS OF A LARGE LANGUAGE MODEL-BASED CHATBOT SERVICE USING RETRIEVAL-AUGMENTED GENERATION ARCHITECTURE (CASE STUDY: UGM ANJEM SERVICE)

Anugrah Dwiki Ardana

22/494995/SV/20942

UGM Anjem is a student community-based ride and delivery service at Universitas Gadjah Mada that has relied solely on WhatsApp as its information channel, requiring repetitive questions to be answered manually by admins. This final project implements a RAG-based chatbot using FastAPI, GPT-4o mini, ChromaDB, and Supabase deployed on the university VPS, and analyzes its service performance using Prometheus and Grafana based on the Four Golden Signals framework. Answer quality was evaluated through a rating feature (5 UGM student respondents) and RAGAS (30 test questions), while performance testing used Locust with four scenarios: 10 users, 25 users, 50 users, and 100 users. Results show a helpful ratio of 92% with an average score of 3.58/4.00. RAGAS yields faithfulness 0.748, answer relevancy 0.693, and answer correctness 0.639, where the high faithfulness score indicates RAG effectively suppresses hallucination. In terms of performance, the system operates reliably under the 10-user and 25-user scenarios with a 0% error rate and stable P50 latency at 3.10 seconds. Performance degradation is detected at the 50-user scenario (error rate 5.22%) and worsens at the 100-user scenario (error rate 28.51%), with the primary bottleneck at the Generation stage via OpenAI API, while local infrastructure saturation remains below 3% CPU and 10% memory across all scenarios.

Keywords: *chatbot, RAG, GPT-4o mini, service performance analysis, RAGAS, load testing.*