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/

DAFTAR ISI

HALAMAN JUDUL i
LEMBAR PENGESAHAN ii
PERNYATAAN BEBAS PLAGIASI iii
KATA PENGANTAR iv
DAFTAR ISI vii
DAFTAR GAMBAR x
DAFTAR TABEL xii
DAFTAR SIMBOL xiii
INTISARI xv
ABSTRACT xvi
BAB I PENDAHULUAN 17
1.1 Latar Belakang 17
1.2 Rumusan Masalah 20
1.3 Tujuan Proyek Akhir 20
1.4 Manfaat Proyek Akhir 21
1.5 Batasan Penelitian 21
1.6 Sistematika Penulisan 22
BAB II KAJIAN PUSTAKA 24
2.1 Tinjauan Pustaka 24
2.2 Dasar Teori 31
2.2.1 *Chatbot* 31
2.2.2 *Large Language Model* (LLM) 32
2.2.3 *Retrieval-Augmented Generation* (RAG) 33
2.2.4 Basis Pengetahuan, *Embedding*, dan Pencarian Vektor 34
2.2.5 *Four Golden Signals* (Google SRE) 35
2.2.6 Prometheus 37
2.2.7 Grafana 38
2.2.8 *Virtual Private Server* dan *Containerization* 39
2.2.9 Penilaian Kepuasan Pengguna 41

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/

2.2.10 RAGAS 41
2.2.11 Load Testing 43
2.2.12 Locust 43
2.2.13 Hipotesis 44
BAB III METODE PENELITIAN 46
3.1 Alat dan Bahan 46
3.2 Tahapan Proyek Akhir 49
3.2.1 Identifikasi Masalah 50
3.2.2 Studi Literatur 51
3.2.3 Perancangan Sistem 51
3.2.4 Implementasi Sistem 52
3.2.5 Pengujian dan Analisis 52
3.3 Desain Arsitektur Sistem 53
3.4 Implementasi Sistem 55
3.4.1 Persiapan Basis Pengetahuan 55
3.4.2 Integrasi LLM 57
3.4.3 Konfigurasi VPS 59
3.4.4 Instalasi Dependensi *Backend* 60
3.4.5 Instalasi Dependensi *Frontend* 61
3.4.6 Proses *Embedding* dan *Vector Store* 62
3.4.7 Implementasi *Pipeline* RAG 63
3.4.8 Konfigurasi *Reverse Proxy* Nginx 67
3.4.9 Kontainerisasi dengan Docker Compose 68
3.4.10 Implementasi Instrumentasi Metrik Prometheus 70
3.4.11 Implementasi Fitur Penilaian Jawaban 73
3.5 Pengujian Sistem 75
3.5.1 Pengujian Kepuasan Pengguna 75

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/

3.5.2 Pengujian Akurasi Jawaban 76
3.5.3 Pengujian Kinerja Layanan 78
BAB IV HASIL DAN PEMBAHAN 80
4.1 Hasil Implementasi Sistem 80
4.1.1 Tampilan Antarmuka *Chatbot* 80
4.1.2 Tampilan Dashboard Grafana 80
4.2 Hasil Kepuasan Pengguna 82
4.2.1 Distribusi Penilaian *Rating* dan Metrik Kepuasan 85
4.2.2 Analisis Skor per Pertanyaan per Responden 88
4.2.3 Pembahasan 89
4.3 Hasil Pengujian Akurasi Jawaban (RAGAS) 90
4.3.1 Hasil Skor Keseluruhan per Metrik 90
4.3.2 Hasil Skor per Kategori Pertanyaan 90
4.3.3 Analisis Pertanyaan dengan Skor Rendah 92
4.3.4 Pembahasan dan Faktor Penyebab 92
4.4 Hasil Pengujian Kinerja Layanan 93
4.4.1 Sinyal *Latency* 93
4.4.2 Sinyal Traffic 97
4.4.3 Sinyal Errors 99
4.4.4 Sinyal Saturation 100
4.4.5 Pembahasan Umum Kinerja Layanan 101
BAB V PENUTUP 104
5.1 Kesimpulan 104
5.2 Saran 104
DAFTAR PUSTAKA 106