MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Sistem Chatbot Helpdesk WhatsApp Berbasis Large Language Model Gemini dengan Fine-Tuning Menggunakan Retrieval-Augmented Generation dan Qdrant Vector Database (Studi Kasus: Universitas Multi Data Palembang)

[1)]M. Rifqi Virgiansyah
Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Indonesia
E-Mail: mrifqivirgiansyah_2226250075p@mhs.mdp.ac.id

[2)]Muhammad Rizky Pribadi
Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Indonesia
E-Mail: rizky@mdp.ac.id

ABSTRACT
Universitas Multi Data Palembang encounters significant challenges in handling a high volume of student inquiries via WhatsApp, resulting in prolonged response times and limited service availability outside operational hours. This study aims to design and implement an intelligent 24-hour helpdesk chatbot system to enhance academic service efficiency. The research employs the Research and Development (R&D) method. The system architecture integrates the Google Gemini Large Language Model (LLM), with performance optimized through the Retrieval-Augmented Generation (RAG) approach and Qdrant Vector Database. This technique enables the chatbot to access an internal knowledge base constructed from official university documents in real-time, thereby minimizing hallucinations. Performance evaluation using the RAGAS framework demonstrates a substantial improvement in Answer Correctness, achieving a score of 0.89, and a Faithfulness score of 0.98 compared to the model without RAG. Furthermore, User Acceptance Testing (UAT) involving 64 respondents yielded an average score of 4.26 (Very Good Category), indicating that the system is feasible for implementation to support rapid and accurate academic information services.
Keyword : ***Chatbot, Large Language Model, Retrieval-Augmented Generation***

PENDAHULUAN

Kemajuan teknologi informasi dan komunikasi telah mempengaruhi hampir semua bidang kehidupan, termasuk perguruan tinggi, sehingga mendorong institusi untuk mengimplementasikan solusi digital demi peningkatan mutu layanan mereka [1]. Berbagai institusi perguruan tinggi telah memanfaatkan platform sistem informasi digital berbasis web sebagai sarana penyampaian informasi penting terkait kampus, seperti informasi akademik, penerimaan mahasiswa baru (PMB) beasiswa, rincian Uang Kuliah Tunggal (UKT), dan lainnya [2].

Saat ini Universitas Multi Data Palembang telah memiliki portal web resmi yang menyediakan berbagai informasi publik serta layanan online untuk menunjang aktivitas akademik dan administrasi mahasiswa. Meskipun sudah menyediakan berbagai layanan online yang interaktif, kenyataannya seringkali mahasiswa akan tetap menghubungi bagian administrasi via call maupun chat di aplikasi WhatsApp di perangkat mobile untuk bertanya agar memperoleh informasi yang memang membutuhkan rincian serta klarifikasi lebih mendalam. Kemudian, pada periode dengan terjadinya lonjakan antrean chat yang dikirim oleh mahasiswa kepada bagian administrasi, sebagai contohnya yaitu periode penerimaan mahasiswa baru atau pada saat registrasi ulang. Lonjakan permintaan layanan seringkali melebihi kapasitas respons admin sehingga menambah beban kinerja bagian administrasi yang harus membalas setiap chat yang masuk ke bagian administrasi kampus. Selain itu, layanan admin online di WhatsApp juga tidak bersifat 24 jam dikarenakan layanan tersebut mempunyai jam operasional yang biasanya mulai beroperasi dari pukul 08.00 pagi hingga 20.00 malam WIB.

Temuan dari wawancara terstruktur dengan Kepala Bagian Administrasi Akademik (BAA), Keuangan (BAK), dan Penerimaan Mahasiswa Baru (PMB) di Universitas Multi Data Palembang secara konsisten menunjukkan bahwa platform utama yang digunakan mahasiswa untuk bertanya adalah WhatsApp. Kendala terbesar yang dihadapi ketiga unit adalah staf harus menjawab pertanyaan yang sama secara berulang, seperti pertanyaan mengenai jadwal pembayaran dan nomor Virtual Account (di BAK), prosedur pendaftaran dan informasi beasiswa (di PMB), hingga alur administrasi pasca-sidang dan kendala teknis sistem (di BAA), padahal sebagian besar informasi tersebut seringkali sudah tersedia di portal resmi (Simponi) atau pedoman akademik.

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Kondisi ini secara nyata mengurangi beban kerja produktif staf, sebagaimana dikeluhkan oleh BAA dan BAK, serta berpotensi menurunkan mutu pelayanan akibat gaya penyampaian yang berbeda antar admin (disebutkan oleh PMB) dan keterbatasan layanan di luar jam kerja. Oleh karena itu, kebutuhan akan sebuah chatbot helpdesk otomatis untuk menangani pertanyaan berulang tersebut dinilai "sangat penting" atau "cukup perlu/penting" oleh ketiga unit terkait, guna meningkatkan efisiensi dan konsistensi layanan informasi.

Sebagai solusi atas keterbatasan tersebut, implementasi chatbot diusulkan sebagai langkah awal sistem cerdas terpadu di Universitas Multi Data Palembang [3]. Namun, tinjauan terhadap penelitian sebelumnya menunjukkan bahwa pengembangan chatbot akademik mayoritas masih menggunakan pendekatan rule-based (berbasis aturan) atau model klasifikasi sederhana [1] [4]. Pendekatan ini cenderung kaku, sulit menangani variasi bahasa alami pengguna yang kompleks, dan tidak mampu memberikan jawaban faktual dari sumber pengetahuan kampus yang dinamis. Dengan demikian, terdapat kesenjangan penelitian (research gap) untuk mengembangkan sistem chatbot helpdesk yang tidak hanya otomatis, tetapi juga cerdas, fleksibel secara percakapan, dan dapat diandalkan secara faktual berdasarkan data internal institusi.

Untuk mengisi kesenjangan tersebut, penelitian ini mengusulkan arsitektur modern yang memadukan Large Language Model (LLM) dengan teknik Retrieval-Augmented Generation (RAG). Model LLM Gemini dari Google dipilih sebagai inti pemrosesan bahasa [5]. Alasan pemilihan ini didukung oleh temuan penelitian terbaru yang menyoroti keunggulan Gemini dalam memberikan respons yang cepat dan lugas (to the point) terhadap pertanyaan spesifik [6], serta kemampuannya yang terpadu dalam menangani berbagai jenis data (multimodal) seperti teks dan gambar. Dari hasil penelitian tersebut, hasil menunjukkan kinerja tinggi dengan Mean Reciprocal Rank (MRR) sebesar 0,83, Exact Match (EM) sebesar 100% [4].

Meskipun LLM Gemini kuat, ia memiliki keterbatasan mendasar; salah satu kendala utama adalah fenomena halusinasi (hallucination), yaitu kecenderungan model menghasilkan jawaban yang tidak akurat atau tidak berdasar fakta [7]. Selain itu, model generatif memiliki keterbatasan memori kontekstual dan tidak terhubung langsung ke sumber data terkini, seperti informasi spesifik dan dinamis Universitas Multi Data Palembang. Untuk mengatasi tantangan tersebut, penelitian ini mengusulkan pendekatan RAG. RAG adalah pendekatan yang memadukan mekanisme pencarian informasi dengan kemampuan generatif model bahasa [8]. Pendekatan ini memungkinkan chatbot memberikan respons yang lebih tepat, kontekstual, dan kaya informasi—tidak hanya jawaban singkat tetapi juga penjelasan atau data pendukung yang bermanfaat bagi pengguna [9]. RAG adalah arsitektur yang memadukan mekanisme pencarian dokumen eksternal (retriever) dengan proses generatif (generator) dari model bahasa. Dalam arsitektur ini, LLM Gemini bertindak sebagai “otak” (generator) yang merangkai jawaban. Namun, agar jawaban itu faktual, ia membutuhkan “buku” (konteks) yang relevan [7]. Disinilah peran Qdrant. Qdrant berfungsi sebagai penyimpanan (storage) khusus yang sangat efisien untuk embedding (representasi vektor) dari semua dokumen kampus [10]. Peran Qdrant bukan hanya sebagai storage pasif, tetapi sebagai fondasi yang memungkinkan komponen retriever melakukan pencarian semantik berkecepatan tinggi untuk menemukan potongan informasi paling relevan dari knowledge base kampus sebagai konteks bagi Gemini.

Sebagai media interaksi, penelitian ini akan memanfaatkan aplikasi WhatsApp. WhatsApp menyediakan antarmuka pemrograman aplikasi atau disingkat API khusus yang memudahkan pengembang dalam membuat chatbot [11]. Pemilihan platform WhatsApp sebagai jalur interaksi utama dipertimbangkan karena tingginya tingkat adopsi aplikasi ini di kalangan mahasiswa serta kemampuannya dalam mendukung komunikasi dua arah secara instan. WhatsApp menjadi salah satu media sosial yang paling populer yang digunakan oleh masyarakat Indonesia [12]. Sepanjang tahun 2024, sekitar 90,9% pengguna internet di Indonesia menggunakan WhatsApp sebagai salah satu platform komunikasi utama mereka [13].

Melalui integrasi chatbot LLM Gemini dengan fine-tuning RAG dan Qdrant ke platform WhatsApp, mahasiswa dapat memperoleh informasi akademik secara real-time dan instan (24 jam). Penelitian ini akan menilai efektivitas solusi melalui metrik kecepatan respons, akurasi jawaban, dan tingkat kepuasan mahasiswa, guna merumuskan model best practice layanan akademik cerdas berbasis chatbot di lingkungan perguruan tinggi Indonesia.

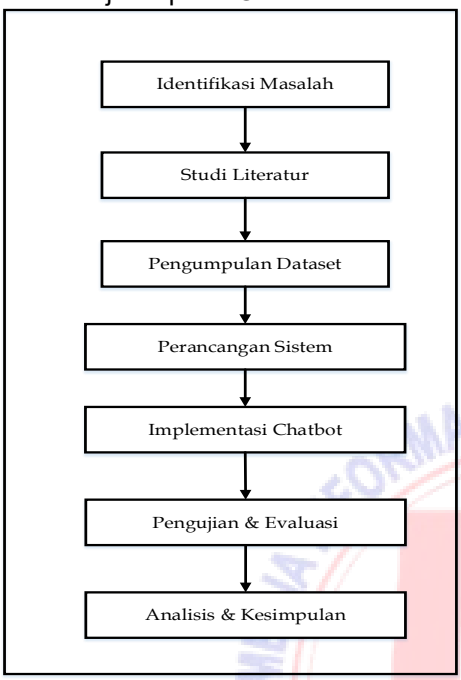
BAHAN DAN METODE

Penelitian ini menerapkan metode Research and Development (R&D) [14], dengan tujuan mengembangkan produk perangkat lunak berupa asisten virtual (chatbot) yang dioptimalkan untuk kebutuhan spesifik institusi. Pendekatan ini dipilih untuk memastikan bahwa sistem yang dibangun mampu melakukan penyesuain domain (fine-tuning) terhadap pengetahuan LLM standar melalui mekanisme injeksi konteks, sehingga efektif dalam menyelesaikan permasalahan operasional layanan akademik di Universitas Multi Data Palembang.

Kerangka kerja pengembangan sistem mengacu pada model pengembangan perangkat lunak linear sekuensial (waterfall model) yang

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

diadaptasi menjadi tujuh tahapan sistematis. Fokus utama pengembangan terletak pada integrasi antarmuka pesan instan WhatsApp dengan kecerdasan buatan generatif (Generative AI) Google Gemini, yang diperkuat dengan teknik RAG dan basis data vektor Qdrant untuk menjamin akurasi informasi lokal. Secara skematis, tahapan metodologi penelitian yang dilakukan disajikan pada Gambar 1.



Gambar 1. Tahapan metodologi penelitian

Berdasarkan Gambar 1 di atas, alur penelitian diawali dengan identifikasi masalah pada layanan helpdesk konvensional, dilanjutkan dengan studi literatur mengenai teknologi LLM dan RAG. Tahap selanjutnya adalah pengumpulan dataset dokumen akademik untuk membangun basis pengetahuan (knowledge base), perancangan sistem yang mengintegrasikan Qdrant dan Gemini, implementasi Chatbot pada platform WhatsApp, pengujian dan evaluasi kinerja menggunakan metrik RAGAS (Retrieval-Augmented Generation Assessment), serta diakhiri dengan analisis hasil dan penarikan kesimpulan mengenai efektivitas sistem dalam konteks studi kasus di Universitas Multi Data Palembang.

Pengumpulan Data dan Preprocessing

Kualitas respon yang dihasilkan oleh sistem RAG sangat bergantung pada kualitas data yang tersimpan dalam knowledge base. Dalam penelitian ini, pengumpulan data dilakukan melalui dua pendekatan, yaitu pengumpulan data primer dan data sekunder.

Pengumpulan Data (Data Collection)

Data yang dihimpun diklasifikasikan berdasarkan sumber pemerolehannya sebagai berikut:

a. Data Primer

Data primer diperoleh melalui wawancara langsung dengan unit-unit administrasi terkait di Universitas Multi Data Palembang, seperti Bagian Administrasi Akademik (BAA), Bagian Administrasi Keuangan (BAK), dan Bagian Penerimaan Mahasiswa Baru (PMB). Wawancara dilakukan untuk memverifikasi prosedur standar operasional (SOP) yang sering ditanyakan, namun informasinya belum termuat secara rinci di dokumen publik. Hasil wawancara ini kemudian didokumentasikan dalam format teks terformat (markdown) untuk menjaga akurasi informasi yang sensitif.

b. Data Sekunder

Data sekunder bersumber dari dokumen digital resmi yang bersifat publik. Data ini dikumpulkan dalam format file yang heterogen untuk melengkapi knowledge base sistem, meliputi:

1) Dokumen Portable (.PDF): Mencakup buku pedoman akademik Universitas Multi Data Palembang, kalender akademik TA 2025/2026, brosur digital pendaftaran S1 dan S2, kode etik mahasiswa, peraturan norma kemahasiswaan, dan lainnya. Dokumen ini merupakan data tidak terstruktur (unstructured data) yang menjadi sumber utama regulasi kampus.
2) Data Web Terstruktur (.JSON): Merupakan hasil web scraping dari sistus resmi mdp.ac.id, pmb.mdp.ac.id, dan kemahasiswaan.mdp.ac.id. Format JSON digunakan untuk menyimpan data hierarkis yang dinamis, seperti daftar biaya kuliah, data dosen, berita kegiatan kampus terbaru, serta artikel lainnya.
3) Dokumen Teks Manual (.MD): Format markdown digunakan untuk menyusun data hasil wawancara (data primer) dan informasi statis seperti struktur organisasi dan FAQ (Frequently Asked Questions) agar mudah dibaca oleh model bahasa.

Pra-pemrosesan Data (Data Preprocessing)

Mengingat variasi format data yang digunakan (PDF, JSON, MD), tahapan preprocessing data menjadi langkah fundamental agar informasi dapat dikonversi menjadi representasi vektor yang akurat. Proses ini diawali dengan tahap ekstraksi dan pembersihan teks (Text Extraction & Cleaning), yaitu sistem melakukan parsing terhadap file PDF dan JSON untuk memisahkan teks murni dari elemen noise. Pembersihan dilakukan secara agresif dengan menghapus karakter non-standar, simbol yang tidak relevan, spasi berlebih, serta elemen tata letak seperti header dan footer yang sering muncul pada setiap halaman dokumen PDF. Setelah teks bersih diperoleh, dilakukan tahap normalisasi teks (Text Normalization) untuk menyeragamkan format, termasuk konversi seluruh karakter menjadi huruf kecil (lowercasing) guna meningkatkan konsistensi pencocokan semantik saat pencarian.

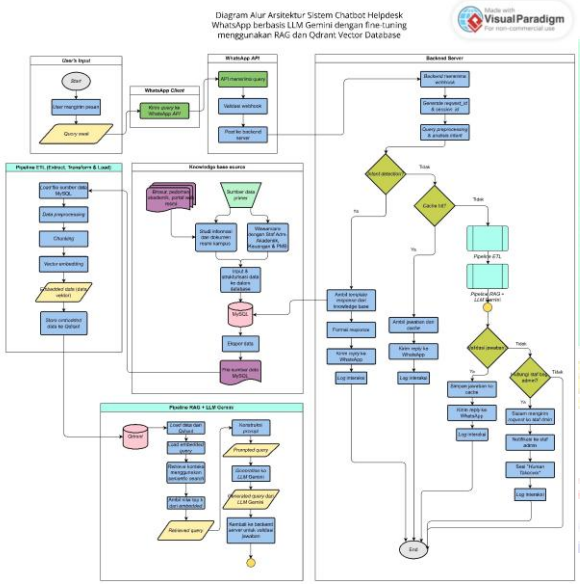
Tahap selanjutnya adalah segmentasi dokumen atau chunking, yaitu proses membagi teks menjadi unit-unit yang lebih kecil berdasarkan struktur sintaksis untuk mempermudah

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

pemahaman dan penyajian informasi oleh model [15]. Penelitian ini memecah dokumen panjang menjadi segmen-segmen (chunks) dengan batasan panjang maksimum 1000 karakter per segmen. Ukuran ini ditetapkan lebih kecil dibandingkan pendekatan umum (2048 karakter) untuk memastikan model embedding dapat menangkap konteks yang sangat spesifik tanpa melebihi batas jendela token (token window), sehingga sistem RAG mampu menemukan jawaban yang presisi pada basis data vektor.

Arsitektur Sistem

Sistem dikembangkan menggunakan arsitektur modular berbasis client-server dengan Node.js sebagai backend orchestrator. Perancangan arsitektur ini bertujuan untuk memastikan skalabilitas dan efisiensi pemrosesan pesan secara real-time. Secara garis besar, alur kerja sistem mengintegrasikan lapisan interaksi pengguna, lapisan logika bisnis, dan lapisan kecerdasan buatan. Ilustrasi lengkap mengenai arsitektur sistem yang diusulkan dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur Sistem

Berdasarkan Gambar 2 di atas, komponen utama penyusun sistem dijelaskan secara rinci sebagai berikut:

1. Antarmuka Pengguna (User Interface Layer):
   Platform WhatsApp dipilih sebagai antarmuka utama karena aksesibilitasnya yang tinggi di banyak kalangan [16], khususnya civitas akademika Universitas Multi Data Palembang, pelajar dan mahasiswa, maupun masyarakat umum. Sistem memanfaatkan library whatsapp-web.js untuk mengemulasikan instans WhatsApp Web, memungkinkan server untuk menerima pesan masuk (incoming messages) dan mengirim respons (outgoing messages) secara programatik (event-driven).
2. Manajemen Sesi dan Logika Routing (Orchestration Layer):
   Modul ini bertindak sebagai otak operasional sistem. Didalamnya terdapat command handler yang menerapkan logika smart routing untuk mengklasifikasikan intensi pesan pengguna. Pesan dikategorikan menjadi dua jenis:
   a) Query Statis
      Sapaan atau pertanyaan umum yang dijawab menggunakan templat baku static content.
   b) Query Dinamis
      Pertanyaan kompleks seputar akademik yang diteruskan ke mesin RAG. Selain itu, modul session manager bertugas menyimpan riwayat percakapan sementara untuk menjaga konteks dialog (multi-turn conversation).
3. Vector Database (Knowledge Base Layer):
   Sistem menggunakan Qdrant sebagai basis data vektor (vector database) untuk penyimpanan memori jangka panjang. Dokumen internal universitas yang telah melalui proses chunking dikonversi menjadi vektor numerik berdimensi tinggi (768 dimensi) menggunakan model embedding text-embedding-004. Qdrant bertugas melakukan pencarian semantik (semantic search) untuk menemukan potongan informasi yang paling relevan dengan query pengguna.
4. LLM Service (Cognitive Layer):
   Google Gemini 2.5 Flash diimplementasikan sebagai generator jawaban cerdas. Model ini dipilih karena kemampuan penalaran (reasoning) yang tinggi dengan latency rendah. LLM bertugas menerima prompt yang telah diperkaya dengan konteks dari Qdrant, lalu mensintesis jawaban yang koheren, dan sesuai dengan gaya bahasa institusi akademik.

Implementasi RAG

Mekanisme RAG diterapkan sebagai solusi inti untuk mengatasi keterbatasan knowledge cutoff dan risiko halusinasi pada LLM. Dalam penelitian ini, alur kerja RAG dirancang melalui tiga tahapan komputasi utama:

1. Semantic Retrieval (pencarian semantik):
   Proses dimulai ketika pengguna mengirimkan query melalui WhatsApp. Sistem melakukan akuisisi teks dan memprosesnya melalui model embedding teks 004. Model ini mengonversi query bahasa alami menjadi representasi vektor numerik berdimensi tinggi. Vektor query tersebut kemudian diproyeksikan ke dalam ruang vektor (vector space) pada database Qdrant.
2. Context Augmentation (augmentasi konteks):
   Tahap ini berfungsi sebagai jembatan antara mesin pencari dan model generatif. Dokumen-dokumen relevan yang diperoleh dari tahap retrieval tidak langsung dikirim ke LLM, melainkan melalui proses augmentasi.

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Sistem menerapkan logika smart routing yang menggabungkan dua sumber data:

a) Data Vektor Dinamis:
Informasi tidak terstruktur dari dokumen PDF (pedoman akademik, brosur PMB, dsb).

b) Data Statis Terstruktur:
Informasi krusial yang membutuhkan presisi absolut, seperti daftar pejabat struktural, rincian biaya kuliah, dsb disuntikkan secara manual (static injection) untuk melengkapi hasil pencarian vektor.

Seluruh data tersebut disusun menjadi satu kesatuan konteks (context window) yang koheren.

3. Generative Response (pembangkitan jawaban):
Pada tahap akhir, konteks yang telah diaugmentasi disuntikkan ke dalam system prompt model Gemini 2.5 Flash. Teknik prompt engineering diterapkan dengan memberikan instruksi ketat (strict instructions) kepada model untuk bertindak sebagai asisten akademik dan hanya menjawab pertanyaan berdasarkan fakta yang tersedia dalam konteks yang diberikan. Hal ini bertujuan untuk memitigasi risiko fabrikasi informasi (hallucination) dan memastikan jawaban yang dihasilkan tetap selaras dengan kebijakan Universitas Multi Data Palembang.

Skenario Pengujian dan Evaluasi

Tahap pengujian dilakukan untuk memverifikasi fungsionalitas sistem dan mengukur kualitas jawaban yang dihasilkan. Pengujian dibagi menjadi dua metode utama:

1. Black Box Testing:
Metode ini digunakan untuk menguji fungsionalitas fitur dasar chatbot tanpa melihat struktur kode internal. Skenario pengujian meliputi validasi respons terhadap salam (greeting), ketepatan penanganan menu bantuan, logika routing pesan (apakah pertanyaan umum dan spesifik ditangani oleh handler yang benar), serta stabilitas koneksi WhatsApp saat menerima beban pesan beruntun.
2. Evaluasi Kinerja Model (RAGAS):
Untuk mengukur efektivitas penerapan RAG secara kuantitatif, penelitian ini menggunakan kerangka kerja RAGAS [3]. Evaluasi dilakukan dengan membandingkan kinerja sistem dalam dua skenario: With RAG (menggunakan dokumen internal) dan No RAG (hanya mengandalkan pengetahuan bawaan LLM). Penilaian dilakukan secara otomatis menggunakan LLM sebagai juri (LLM-as-a-Judge) terhadap lima metrik utama, yaitu:
   a) Faithfulness (kesetiaan):
   Mengukur sejauh mana jawaban yang dihasilkan diturunkan murni dari konteks dokumen yang diambil, tanpa adanya halusinasi atau penambahan informasi yang tidak berdasar.
   b) Answer Relevance (relevansi jawaban):
   Mengukur seberapa relevan jawaban yang diberikan terhadap pertanyaan pengguna, memastikan jawaban tidak menyimpang dari topik.
   c) Context Precision (presisi konteks):
   Mengukur apakah dokumen-dokumen yang diambil oleh sistem pencarian (retriever) di urutan teratas benar-benar relevan dengan pertanyaan.
   d) Context Recall (kelengkapan konteks):
   Mengukur apakah sistem pencarian berhasil mengambil seluruh informasi yang diperlukan dari knowledge base untuk menjawab pertanyaan secara lengkap sesuai dengan kebenaran dasar (ground truth).
   e) Answer Correctness (kebenaran jawaban):
   Mengukur akurasi faktual dan semantik dari jawaban chatbot dibandingkan dengan kunci jawaban referensi (ground truth) yang telah disiapkan sebelumnya.

HASIL DAN PEMBAHASAN

1. Implementasi Sistem

Tahap implementasi merupakan realisasi dari rancangan arsitektur yang telah didefinisikan sebelumnya. Sistem chatbot helpdesk ini dibangun diatas lingkungan eksekusi Node.js yang bertindak sebagai backend service. Layanan ini berhasil mengorkestrasi komunikasi dua arah antara antarmuka pengguna (WhatsApp) dengan lapisan kecerdasan buatan (Google Gemini API) dan memori jangka panjang (Qdrant Vector Database) secara real-time.

a) Implementasi Antarmuka dan Manajemen Sesi

Antarmuka pengguna diimplementasikan menggunakan platform WhatsApp melalui integrasi library whatsapp-web.js. Pendekatan ini memungkinkan pengguna untuk berinteraksi dengan sistem tanpa perlu mengunduh aplikasi tambahan atau melakukan autentikasi ulang.

Pada saat pengguna mengirimkan pesan pertama kali, sistem secara otomatis menginisialisasi sesi baru dan mengirimkan respons sapaan (greeting) yang ramah serta informatif. Hal ini bertujuan untuk memberikan panduan navigasi awal kepada pengguna mengenai topik apa saja yang dapat ditanyakan. Visualisasi tampilan antarmuka saat sistem menangani interaksi awal dan memberikan menu bantuan dapat dilihat pada Gambar 3.

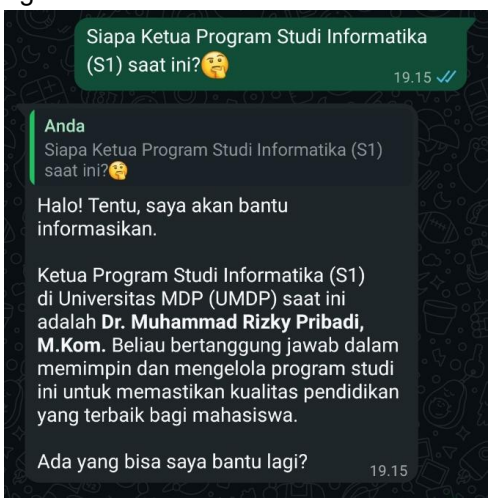
MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Gambar 3. Tampilan respon awal dan menu bantuan chatbot

b) Implementasi Mekanisme RAG

Kemampuan inti sistem dalam menangani pertanyaan akademik yang kompleks dibuktikan melalui integrasi RAG. Ketika sistem mendeteksi pertanyaan spesifik (seperti prosedur administrasi atau rincian biaya), modul command handler meneruskan query tersebut ke mesin pencari vektor.

Sistem kemudian mengambil konteks relevan dari dokumen internal universitas dan menyusun jawaban yang presisi. Sebagaimana ditunjukkan pada Gambar 4, chatbot mampu memberikan penjelasan terperinci mengenai prosedur akademik (seperti pengurusan KTM atau biaya kuliah) beserta langkah-langkah solutifnya. Jawaban ini dihasilkan bukan berdasarkan pengetahuan umum model semata, melainkan disintesis dari fakta dokumen yang ditemukan oleh sistem.



Gambar 4. Tampilan jawaban *chatbot* menggunakan konteks RAG

Implementasi ini memvalidasi bahwa alur data dari query pengugna hingga generated response berjalan dengan latency yang rendah dan format jawaban yang terstruktur, sesuai dengan standar layanan akademik Universitas Multi Data Palembang.

2. Pengujian Kinerja Teknis (Evaluasi RAGAS)

Evaluasi teknis bertujuan untuk mengukur kualitas jawaban chatbot secara objektif menggunakan kerangka kerja RAGAS. Pengujian ini tidak hanya berfokus pada kebenaran jawaban semata, tetapi juga melibatkan lima aspek metrik utama untuk memastikan sistem bekerja sesuai standar akademik.

Pengujian dilakukan dengan membandingkan dua skenario: Skenario A (No RAG), yaitu chatbot hanya mengandalkan pengetahuan bawaan model, dan Skenario B (With RAG), yaitu chatbot dibantu oleh dokumen internal universitas.

Hasil pengukuran terhadap lima metrik RAGAS disajikan secara komprehensif pada Tabel 1.

Tabel 1. Hasil Evaluasi Metrik RAGAS (Skala 0.0 – 1.0)

Metrik Evaluasi	Deskripsi Metrik	*No* RAG	*With* RAG	Δ
Faithfulness	Tingkat konsistensi jawaban terhadap konteks (anti-halusinasi)	0.00	0.98	+0.98
Answer Relevance	Relevansi jawaban terhadap pertanyaan pengguna	0.20	1.00	+0.80
Context Precision	Ketepatan sistem dalam mengambil dokumen yang relevan di urutan teratas	0.00	0.60	+0.60
Context Recall	Kelengkapan informasi yang berhasil diambil dari database	0.00	0.72	+0.72
Answer Correctness	Akurasi fakta jawaban dibandingkan dengan ground truth	0.08	0.89	+0.81

Berdasarkan Tabel 1, dapat dianalisis bahwa:

a. Peningkatan Akurasi (Correctness): Integrasi RAG meningkatkan akurasi dari 0.08 menjadi 0.89. Ini membuktikan bahwa tanpa RAG, model LLM mengalami kegagalan total dalam menjawab pertanyaan spesifik institusi.

b. Eliminasi Halusinasi (Faithfulness): Skor faithfulness mencapai 0.98, yang berarti hampir seluruh jawaban yang dihasilkan oleh chatbot didasarkan pada dokumen resmi universitas, bukan karangan model.

c. Kualitas Retrieval: Skor context precision (0.60) dan recall (0.72) menunjukkan bahwa mesin pencari vektor Qdrant cukup efektif dalam menemukan dokumen yang relevan,

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

meskipun masih terdapat ruang optimasi pada strategi chunking untuk meningkatkan presisi dokumen teratas.

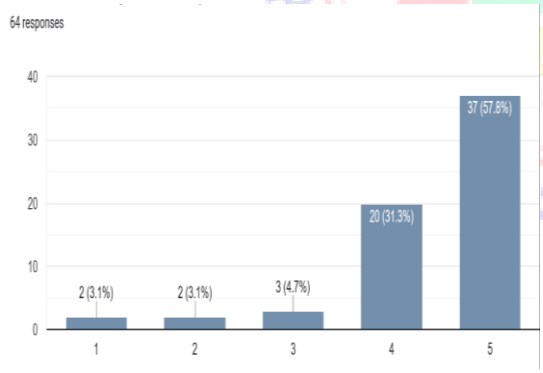
3. Pengujian Akseptansi Pengguna (User Acceptance Testing)

Selain evaluasi teknis, penelitian ini juga melakukan uji validasi empiris kepada pengguna akhir untuk mengukur tingkat kepuasan, kegunaan, dan akurasi informasi yang dirasakan pengguna. User Acceptance Testing (UAT) merupakan mekanisme validasi akhir yang dilakukan langsung oleh pengguna untuk menentukan kelayakan operasional sistem [17]. Pengujian dilakukan dengan menyebarkan kuesioner di Google Form kepada 64 responden yang terdiri dari kelompok heterogen, meliputi civitas akademika Universitas Multi Data Palembang hingga masyarakat umum.

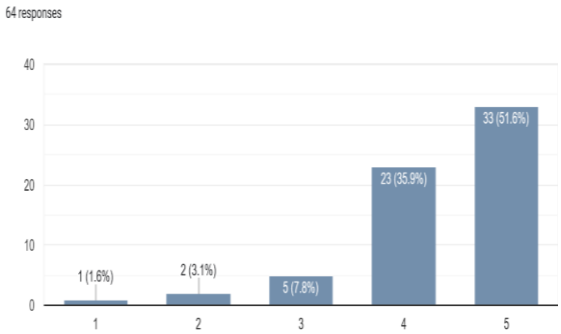
Instrumen kuesioner dirancang menggunakan skala Likert 5 point, yaitu (1) Sangat Tidak Setuju, (2) Tidak Setuju, (3) Netral, (4) Setuju, dan (5) Sangat Setuju. Evaluasi dibagi menjadi dua dimensi utama, yaitu user experience (pengalaman pengguna) dan kualitas jawaban chatbot.

a. Analisis Pengalaman Pengguna (User Experience)

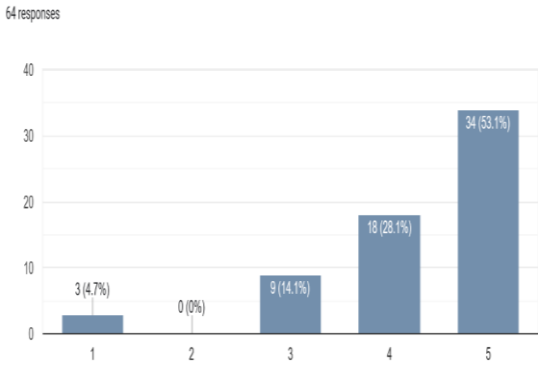
Dimensi ini mengukur interaksi antarmuka dan kemudahan pengguna sistem. Berdasarkan rekapitulasi data kuesioner, visualisasi distribusi jawaban responden untuk setiap indikator pengalaman pengguna dapat dilihat pada Gambar 5 sampai Gambar 8.



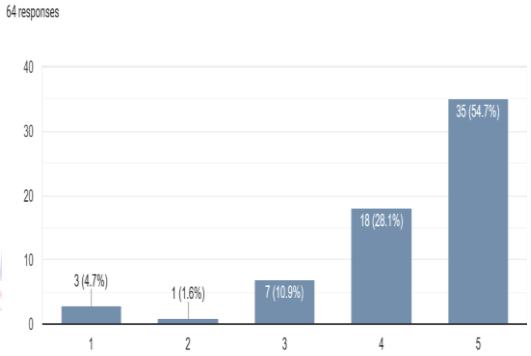
Gambar 5. Grafik distribusi indikator kemudahan penggunaan



Gambar 6. Grafik distribusi indikator gaya bahasa



Gambar 7. Grafik distribusi indikator kecepatan respons



Gambar 8. Grafik distribusi indikator pemahaman konteks

Berdasarkan grafik diatas, rincian persentase jawaban responden dan nilai rata-rata (mean) untuk setiap indikator disajikan secara rinci pada Tabel 2.

Tabel 2. Hasil evaluasi pengalaman pengguna (user experience)

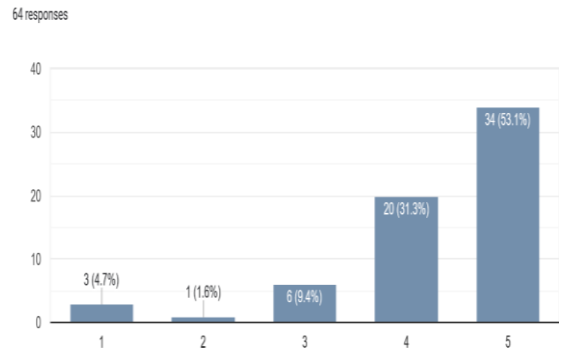
No	Indikator Penilaian	STS (1)	TS (2)	N (3)	S (4)	SS (5)	*Mean*	Kategori
1.	Kemudahan penggunaan	3.1%	3.1%	4.7%	31.3%	57.8%	4.38	Sangat Baik
2.	Gaya bahasa	1.6%	3.1%	7.8%	35.9%	51.6%	4.33	Sangat Baik
3.	Kecepatan respons	4.7%	0%	14.1%	28.1%	53.1%	4.30	Sangat Baik
4.	Pemahaman konteks	4.7%	1.6%	10.9%	28.1%	54.7%	4.30	Sangat Baik
	Rata-Rata Keseluruhan						4.32	**Sangat Baik**

Tabel 2 menunjukkan bahwa aspek kemudahan penggunaan dan kecepatan respons mendominasi kategori ”Sangat Setuju” (diatas 50%). Hal ini mengindikasikan bahwa penggunaan platform WhatsApp efektif menurunkan hambatan adaptasi pengguna, serta sistem berhasil memberikan respons real-time yang mengatasi masalah antrean layanan manual.

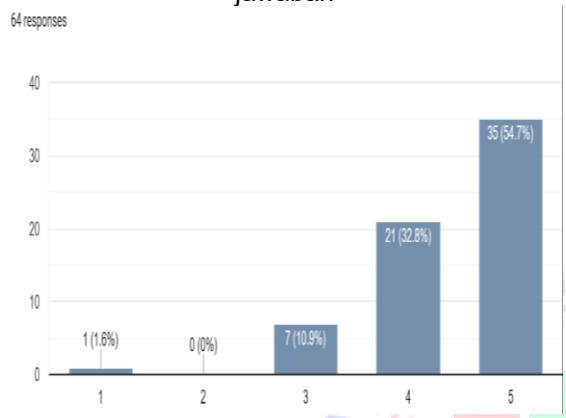
b. Analisis Kualitas Jawaban (Chatbot Quality)

Dimensi ini berfokus pada akurasi dan relevansi konten yang dihasilkan oleh mesin RAG. Visualisasi distribusi jawaban responden untuk lima indikator kualitas jawaban disajikan pada Gambar 9 sampai Gambar 13.

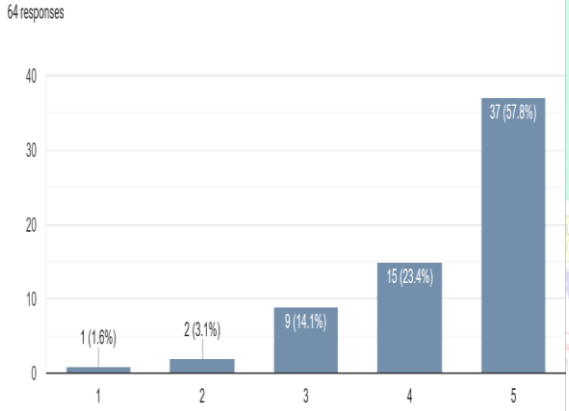
MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089



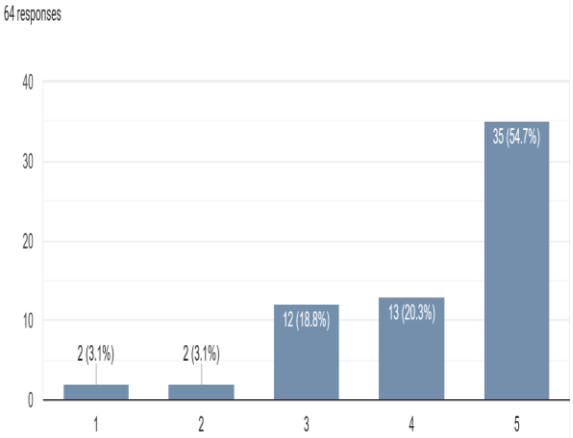
Gambar 9. Grafik distribusi indikator relevansi jawaban



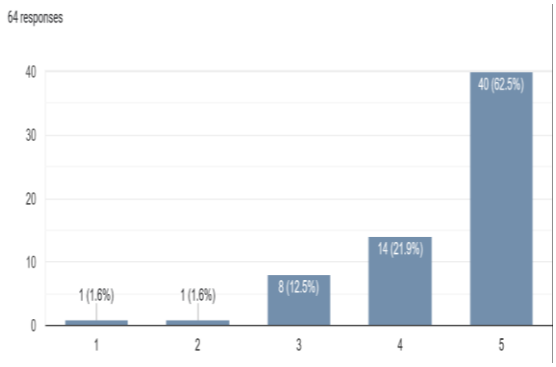
Gambar 10. Grafik distribusi indikator akurasi informasi



Gambar 11. Grafik distribusi indikator kejelasan informasi



Gambar 12. Grafik distribusi indikator kemampuan menjawab spesifik



Gambar 13. Grafik distribusi indikator kepuasan pengguna

Rincian persentase jawaban responden terhadap kelima indikator kualitas jawaban chatbot dirangkum dalam Tabel 3.

Tabel 3. Hasil evaluasi kualitas jawaban chatbot

No	Indikator Penilaian	STS (1)	TS (2)	N (3)	S (4)	SS (5)	Mean	Kategori
1.	Relevansi jawaban	4.7%	1.6%	9.4%	31.3%	53.1%	4.27	Sangat Baik
2.	Akurasi informasi	1.6%	0%	10.9%	32.8%	54.7%	4.23	Sangat Baik
3.	Kejelasan informasi	1.6%	3.1%	14.1%	23.4%	57.8%	4.08	Baik
4.	Jawaban spesifik	3.1%	3.1%	18.8%	20.3%	54.7%	4.20	Sangat Baik
5.	Kepuasan pengguna	1.6%	1.6%	12.5%	21.9%	62.5%	4.28	Sangat Baik
	Rata-Rata Keseluruhan						4.21	**Sangat Baik**

Berdasarkan Tabel 3, mayoritas responden memberikan respon positif terhadap relevansi dan akurasi jawaban dengan capaian skor rata-rata diatas 4.0. Temuan ini memvalidasi hasil evaluasi teknik RAGAS sebelumnya, yang membuktikan bahwa integrasi dokumen internal melalui Qdrant terbukti mampu menyajikan informasi yang faktual. Lebih lanjut, indikator kepuasan pengguna yang mencapai skor 4.13 mengindikasikan bahwa sistem chatbot ini layak (feasible) untuk diterapkan sebagai pendamping layanan helpdesk utama di Universitas Multi Data Palembang.

4. Analisis Kasus dan Pembahasan

Keberhasilan sistem tidak hanya bergantung pada akurasi RAG, tetapi juga pada logika routing pesan yang diterapkan. Berdasarkan analisis log pengujian, terdapat dua temuan penting:

a. Penanganan pertanyaan spesifik: Pada kasus pertanyaan “Siapa Ketua Program Studi Informatika?”, sistem menerapkan logika Smart Routing yang mendeteksi intensi pengguna mencari nama pejabat. Sistem secara otomatis menyuntikkan data terstruktur dari modul statisk e dalam prompt Gemini. Hasilnya, chatbot mampu menyebutkan nama ”Dr. Muhammad Rizky Pribadi, M.Kom.” dengan presisi 100%, mengatasi kelemahan umum pencarian vektor yang terkadang kurang akurat dalam membedakan entitas nama orang.
b. Penanganan istilah khusus: Pada skenario PMB, pengguna menanyakan tentang ”Jalur

Early Bird". Meskipun istilah ini merupakan terminologi pemasaran spesifik, mekanisme RAG berhasil mengambil potongan dokumen (chunks) yang relevan dari brosur PMB yang telah diindeks ke Qdrant. LLM kemudian mensintesis informasi tersebut menjadi jawaban yang koheren, menjelaskan detail potongan biaya yang didapat mahasiswa.

Secara keseluruhan, integrasi antara Gemini 2.5 Flash dan Qdrant terbukti mampu menjawab tantangan layanan informasi akademik yang membutuhkan akurasi tinggi dan ketersediaan 24 jam.

KESIMPULAN

Penelitian ini berhasil merancang dan mengimplementasikan sistem chatbot helpdesk akademik berbasis WhatsApp menggunakan pendekatan RAG dengan model Gemini 2.5 Flash dan basis data vektor Qdrant. Berdasarkan hasil pengujian, penerapan teknik RAG terbukti efektif mengatasi kelemahan halusinasi pada LLM standar, yang ditunjukkan oleh peningkatan signifikan skor Answer Correctness dari 0.22 (No RAG) menjadi 0.89 (With RAG). Selain itu, sistem mencapai tingkat Faithfulness sebesar 0.98, yang mengindikasikan bahwa penyuntikan konteks dokumen internal sangat krusial untuk menghasilkan jawaban yang konsisten dengan dokumen resmi universitas serta meminimalisir risiko penyebaran informasi yang menyesatkan. Dari sisi arsitektur sistem, implementasi logika smart routing yang memadukan pencarian vektor dengan injeksi data terstruktur terbukti andal dalam menjawab pertanyaan yang membutuhkan presisi tinggi, seperti data pejabat struktural dan rincian biaya kuliah, yang sebelumnya sulit dijawab secara akurat oleh model generatif murni. Secara keseluruhan, sistem ini berhasil memberikan solusi layanan informasi yang tersedia 24 jam secara real-time, sehingga efektif mengatasi kendala antrean pesan yang panjang dan keterbatasan jam operasional staf administrasi manual di Universtias Multi Data Palembang.

Untuk pengembangan penelitian selanjutnya, disarankan agar sistem dapat diintegrasikan dengan teknologi Optical Character Recognition (OCR) agar chatbot mampu membaca dan menjelaskan informasi yang terdapat dalam brosur berbentuk gambar. Selain itu, pengembangan dashboard analitik berbasis web diperlukan untuk membantu manajemen universitas memantau tren pertanyaan dan keluhan mahasiswa secara visual.

DAFTAR PUSTAKA

[1] D. M. Alfiansyah, W. Wiilys, L. Setiyani, D. F. Wati, and D. Dedih, "Pengembangan Chatbot Berbasis Web untuk Layanan Informasi di Horizon University," bit-Tech, vol. 7, no. 3, pp. 1068–1077, Apr. 2025, doi: 10.32877/bt.v7i3.2318.

[2] A. Z. Amrullah, A. S. Anas, and Primajati. Gilang, "Implementasi Chatbot sebagai Virtual Assistant Penerimaan Mahasiswa Baru pada Universitas Bumigora," Jurnal Bumigora Information Technology (BITe), vol. 4, no. 1, pp. 17–26, Jun. 2022, doi: 10.30812/bite.v4i1.1664.

[3] Y. Tribber, K. Kusnadi, and M. Asfi, "Implementasi Retrieval Augmented Generation untuk Layanan Informasi Kampus dengan Chatbot Berbasis AI," Prosiding SISFOTEK, vol. 8, no. 1, pp. 594–600, Nov. 2024, Accessed: Jun. 25, 2025. [Online]. Available: https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/560

[4] T. Q. Ramadhani, N. Q. Nada, and N. D. S, "Penerapan Metode Retrieval-Augmented Generation (RAG) Pada Chatbot E-Commerce Berbasis Gemini Ai," Jurnal Ilmiah ILKOMINFO - Ilmu Komputer & Informatika, vol. 8, no. 2, pp. 301–313, Jul. 2025, doi: 10.47324/ilkominfo.v8i2.384.

[5] N. Rachmat and D. P. Kesuma, "Implementasi LLM Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android," Jurnal Ilmu Komputer (JUIK), vol. 4, no. 1, p. 40, Feb. 2024, doi: 10.31314/juik.v4i1.2831.

[6] A. Basri and E. Ernawati, "Pemanfaatan Chatbot AI Untuk Mendukung Penyelesaian Tugas Akademik Mahasiswa Matematika: Studi Kasus Universitas Muslim Maros," Prosiding Seminar Nasiona FKIP Universitas Muslim Maros, vol. 2, no. 1, pp. 55–60, Jul. 2025, Accessed: Oct. 22, 2025. [Online]. Available: https://ejournals.umma.ac.id/index.php/semnas/article/view/2818

[7] M. D. A. Muhajir, N. Prastiti, and M. Koeshardianto, "IMPLEMENTASI CHATBOT MENGGUNAKAN FRAMEWORK LANGCHAIN BERBASIS LLM GPT (STUDI KASUS : PANDUAN AKADEMIK UNIVERSITAS TRUNOJOYO)," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 2, pp. 2151–2158, Mar. 2025, doi: 10.36040/jati.v9i2.13003.

[8] G. D. Albert and A. Voutama, "PENGEMBANGAN CHATBOT BERBASIS PDF MENGGUNAKAN LOCAL RETRIEVAL-AUGMENTED GENERATION (RAG) DAN OLLAMA," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 2, pp. 937–944, Apr. 2025, doi: 10.23960/jitet.v13i2.6361.

[9] I. I. R. Pratama and B. Sisephaputra, "Pengembangan Sistem Helpdesk

Menggunakan Chatbot Dengan Metode Retrieval-augmented Generation (Rag)," Journal of Informatics and Computer Science (JINACS), vol. 6, no. 3, pp. 696–710, Nov. 2024, doi: 10.26740/jinacs.v6n03.p696-710.

[10] A. A. Sujana, I. Yustiana, and A. Sujjada, "Integration of Qdrant Vector Database and DeepSeek AI for Automated Chatbots on E-Commerce Applications," Jurnal CoSciTech (Computer Science and Information Technology), vol. 6, no. 2, pp. 311–318, Sep. 2025, doi: https://doi.org/10.37859/coscitech.v6i2.9668.

[11] G. P. M. Putra, A. Tenriawaru, and G. Gunawan, "Rancang Bangun Virtual Assistant Chatbot Menggunakan Node.Js pada Layanan Sistem Informasi Akademik," in Prosiding Seminar Nasional Pemanfaatan Sains Dan Teknologi Informasi, Kolaka: Universitas Sembilanbelas November Kolaka, Nov. 2023, pp. 1–5. Accessed: Jun. 25, 2025. [Online]. Available: https://epublikasi.digitallinnovation.com/index.php/sempatin/article/view/47

[12] F. P. N. Koten, A. Jufriansah, and H. Hikmatiar, "Analisis Penggunaan Aplikasi Whatsapp sebagai Media Informasi dalam Pembelajaran: Literature Review," Jurnal Ilmu Pendidikan (JIP) STKIP Kusuma Negara, vol. 14, no. 1, pp. 72–84, Jul. 2022, doi: 10.37640/jip.v14i1.1409.

[13] DataReportal, "Digital 2024: Indonesia." Accessed: May 24, 2025. [Online]. Available: https://datareportal.com/reports/digital-2024-indonesia

[14] L. R. Hidayat, I. G. P. S. Wijaya, and R. Dwiyansaputra, "OPTIMALISASI LAYANAN SISTEM INFORMASI MAHASISWA DENGAN INTEGRASI TELEGRAM : CHATBOT RETRIEVAL-AUGMENTED-GENERATION BERBASIS LARGE LANGUAGE MODEL," Jurnal Teknologi Informasi, Komputer, dan Aplikasinya (JTIKA), vol. 7, no. 1, pp. 121–131, Mar. 2025, doi: 10.29303/jtika.v7i1.459.

[15] S. A. Talaohu, R. Soekarta, and M. Surahmanto, "Implementasi LLM Pada Chatbot PMB Universitas Muhammadiyah Sorong Menggunakan Metode RAG Berbasis Website," Framework : Jurnal Ilmu Komputer Dan Informatika, vol. 3, no. 2, pp. 1–11, Aug. 2025, doi: https://doi.org/10.33506/framework.v3i02.4790.

[16] S. Khadafi, R. A. Saputra, and R. Uttungga, "Implementasi Chatbot Informasi Akademik Menggunakan Jaro Winkler pada Program Studi Sistem Informasi-ITATS," in Seminar Nasional Sains dan Teknologi Terapan, Surabaya: Institut Teknologi Adhi Tama Surabaya, 2024, pp. 1–11. Accessed: Aug. 30, 2025. [Online]. Available: https://ejurnal.itats.ac.id/sntekpan/article/view/6575/0

[17] G. Guntoro, L. Costaner, and L. Lisnawita, "Aplikasi Chatbot untuk Layanan Informasi dan Akademik Kampus Berbasis Artificial Intelligence Markup Language (AIML)," Digital Zone: Jurnal Teknologi Informasi dan Komunikasi, vol. 11, no. 2, pp. 291–300, Nov. 2020, doi: 10.31849/digitalzone.v11i2.5049.