Jurnal Minfo Polgan (JMP)
Volume 15, Nomor 2, Mei-Agustus 2026
DOI : https://doi.org/10.33395/jmp.v15i2.16268
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Uji Kinerja Model Generative AI Menggunakan Dataset Informasi Kampus Melalui Sistem Otomatisasi n8n

[1]Sylbila Najla Sobia Naim, [2]I Kadek Dwi Gandika Supartha, [3]Ayu Manik Dirgayusari, [4]I Putu Agus Eka Darma Udayana, [5]Indra Pratistha
[1,2,3,4,5]Informatika, Institut Bisnis dan Teknologi Indonesia, Bali, Indonesia
*Korespondensi: sylbilanajla@gmail.com

Submit : 05 Mei 2026 | **Diterima** : 30 Mei 2026 | **Terbit** : 09 Juni 2026

ABSTRACT

The development of Generative Artificial Intelligence (AI) has encouraged the use of Large Language Models (LLMs) in campus information chatbot systems. However, Generative AI still has limitations, such as hallucination, the inability to consistently generate answers based on specific data, and dependency on hardware performance. This study aims to analyze the performance of several Generative AI models in a Retrieval-Augmented Generation (RAG)-based chatbot system. This research used a comparative experimental method with a quantitative approach involving three Generative AI models, namely Qwen3-32b, GPT-OSS 20b, and Llama 3.1 8b. The chatbot system was integrated using n8n, Pinecone Vector Store, Cohere embedding model, and Groq API. The dataset consisted of 45 categories of campus information. Exact Match (EM) and ROUGE evaluations were conducted using 88 test questions, where each model was tested four times using the same retrieval configuration. In addition, User Acceptance Testing (UAT) was conducted involving 11 active students using 30 questions across five evaluation aspects. The results showed that all models produced an EM score of 0% because the answers were generated in paraphrased variations. In the ROUGE evaluation, Llama 3.1 8b achieved the best performance with ROUGE-1 scores of 69%–72% and ROUGE-L scores of 61%–66%. GPT-OSS 20b achieved the highest user satisfaction level and the fastest response time, while Qwen3-32b excelled in answer clarity. Hardware specifications also affected chatbot performance, particularly response time and system efficiency. Based on the results, each model demonstrated different performance characteristics, indicating that model selection can be adjusted according to the implementation needs of RAG-based campus information chatbot systems.

***Keywords**: Chatbot, Comparative experiment, LLM, RAG, User Acceptance Testing*

ABSTRAK

Perkembangan Generative Artificial Intelligence (AI) mendorong pemanfaatan Large Language Model (LLM) pada sistem chatbot informasi kampus. Namun, Generative AI masih memiliki keterbatasan berupa hallucination, ketidakmampuan menghasilkan jawaban berbasis data spesifik secara konsisten, serta dipengaruhi oleh performa perangkat keras yang digunakan. Penelitian ini bertujuan untuk menganalisis performa beberapa model Generative AI pada chatbot berbasis Retrieval-Augmented Generation (RAG). Metode penelitian menggunakan eksperimen komparatif dengan pendekatan kuantitatif terhadap tiga model Generative AI, yaitu Qwen3-32b, GPT-OSS 20b, dan Llama 3.1 8b. Sistem chatbot diintegrasikan menggunakan n8n, Pinecone Vector Store, Cohere embedding model, dan Groq API. Dataset penelitian mencakup 45 kategori informasi kampus. Pengujian Exact Match (EM) dan ROUGE dilakukan menggunakan 88 pertanyaan uji, di mana setiap model diuji sebanyak empat kali menggunakan konfigurasi retrieval yang sama. Selain itu, User Acceptance Test (UAT) dilakukan terhadap 11 mahasiswa aktif menggunakan 30 pertanyaan dalam lima aspek penilaian. Hasil penelitian menunjukkan bahwa seluruh model menghasilkan nilai EM sebesar 0% karena jawaban berbentuk variasi parafrase. Pada pengujian ROUGE, model Llama 3.1 8b memperoleh performa terbaik dengan nilai ROUGE-1 sebesar 69%–72% dan ROUGE-L sebesar 61%–66%. GPT-OSS 20b memperoleh tingkat kepuasan pengguna tertinggi dan waktu respons tercepat, sedangkan Qwen3-32b unggul pada kejelasan jawaban. Spesifikasi perangkat keras juga mempengaruhi performa chatbot, terutama pada waktu respons dan efisiensi sistem. Berdasarkan hasil

penelitian, setiap model memiliki karakteristik performa yang berbeda sehingga pemilihan model dapat disesuaikan dengan kebutuhan implementasi chatbot informasi kampus berbasis RAG.

Kata Kunci: Chatbot, Eksperimen komparatif, LLM, RAG, UAT

PENDAHULUAN

Perkembangan teknologi Artificial Intelligence (AI) atau kecerdasan buatan telah membawa perubahan besar dalam berbagai bidang kehidupan, termasuk pendidikan, bisnis, dan layanan informasi publik. Salah satu perkembangan penting dalam bidang AI adalah munculnya Generative Artificial Intelligence dan Large Language Model (LLM) yang mampu memahami serta menghasilkan teks secara natural menyerupai bahasa manusia (Zubaedah et al., 2026). Generative AI menjadi tren karena kemampuannya dalam menciptakan konten baru secara otomatis berdasarkan data yang telah dipelajari sebelumnya, seperti teks, gambar, video, maupun kode program. Teknologi ini juga mampu meningkatkan efisiensi dan produktivitas pada berbagai sektor, termasuk pendidikan dan layanan informasi digital (Zubaedah et al., 2026). Large Language Model (LLM) merupakan model bahasa yang telah dilatih menggunakan berbagai sumber pengetahuan dalam jumlah besar sehingga mampu memahami konteks dan menghasilkan respons berbasis bahasa alami. Model-model ini dimanfaatkan secara luas untuk mendukung sistem tanya jawab otomatis, asisten virtual, hingga pengelolaan sistem informasi yang membutuhkan interaksi berbasis bahasa (Al-kfairy et al., 2024). Pemanfaatan LLM pada sistem chatbot memungkinkan pengguna memperoleh informasi secara lebih cepat, interaktif, dan natural dibandingkan sistem chatbot konvensional (Romadhona Kusuma & Ternado, 2025).

Dalam konteks perguruan tinggi, penyediaan layanan informasi kampus merupakan hal penting bagi mahasiswa maupun calon mahasiswa (Prasojo et al., 2024).Namun, layanan informasi seperti kunjungan langsung ke kampus, pencarian melalui situs web, atau penggunaan layanan chatting berbasis WhatsApp sering kali dinilai kurang efisien dan memerlukan waktu yang cukup lama untuk memperoleh jawaban yang cepat dan akurat terutama ketika pertanyaan bersifat mendalam dan spesifik terhadap informasi kampus (Prasojo et al., 2024) Penelitian sebelumnya, seperti "Aplikasi Chatbot Berbasis Telegram untuk Layanan Informasi dan Akademik Kampus Universitas Ma'arif Nahdlatul Ulama Kebumen," menunjukkan bahwa sistem berbasis aturan (rule-based) seperti Artificial Intelligence Markup Language (AIML) dapat memberikan tingkat akurasi hingga 88% pada User Acceptance Test (UAT). Akan tetapi, pendekatan tersebut masih memiliki keterbatasan dalam skalabilitas, pemeliharaan data, dan yang utama adalah memahami konteks yang lebih kompleks dan memberikan jawaban yang adaptif terhadap berbagai bentuk pertanyaan (Prasojo et al., 2024). Selain itu, Artificial Intelligence Markup Language (AIML) punya keterbatasan tentang sifatnya yang berbasis aturan, sehingga menghambat kemampuan model dalam memproses dan merespons pertanyaan yang ambigu, kompleks, atau menuntut pemahaman kontekstual di luar kerangka pemrograman awalnya (Attigeri, Agrawal, & Kolekar, 2024).

Keterbatasan tersebut membuka peluang penerapan Generative AI, khususnya Large Language Model (LLM), yang memiliki kemampuan memahami konteks dan menghasilkan jawaban yang lebih dinamis dan adaptif. Model Large Language Model (LLM) menjadi kuat karena besarnya data pelatihan yang digunakan. Namun, ketergantungan pada data tersebut menimbulkan kelemahan berupa keterbatasan pembaruan pengetahuan model. Hal ini menyebabkan Large Language Model (LLM) kurang mampu menyajikan informasi, peristiwa, atau penemuan yang bersifat terkini (Ling et al., 2025). Selain itu, munculnya fenomena hallucination atau keluaran yang tidak sesuai fakta, serta keterbatasan model dalam mengakses pengetahuan eksternal yang spesifik dan terbaru turut menjadi tantangan dalam penerapan Large Language Model (LLM), khususnya pada domain kampus (Lewis et al., 2020). Penelitian mengenai Retrieval-Augmented Generation (RAG) menunjukkan bahwa integrasi antara mekanisme pencarian data eksternal dan model generatif dapat mengurangi kesalahan informasi, meningkatkan keakuratan, serta meningkatkan keandalan sistem tanya jawab (Lewis et al., 2020). Dalam pendekatan RAG, Large Language Model (LLM) dapat memanfaatkan konteks eksternal yang relevan sehingga jawaban yang dihasilkan menjadi lebih sesuai dengan domain spesifik yang digunakan (Ling et al., 2025).

Untuk mengetahui tingkat ketepatan model LLM dalam menghasilkan jawaban berdasarkan dataset domain-spesifik seperti informasi kampus, penelitian ini menerapkan metode evaluasi yang mampu mengukur kualitas jawaban model secara objektif. Oleh karena itu, digunakan dua metrik evaluasi yang telah menjadi standar dalam penelitian Question

Answering (QA), yaitu Exact Match (EM) dan ROUGE (Risch, Möller, Gutsch, & Pietsch, 2021). Exact Match (EM) digunakan untuk menilai apakah jawaban model sama persis dengan jawaban acuan sehingga efektif untuk mengevaluasi pertanyaan faktual yang membutuhkan ketepatan penuh (Risch et al., 2021). Sementara itu, ROUGE digunakan untuk menilai tingkat kemiripan jawaban model dengan jawaban acuan berdasarkan kesamaan kata dan struktur kalimat (Aliphadji Talaohu, Soekarta, & Surahmanto, 2025) Penggunaan kombinasi Exact Match (EM) dan ROUGE memberikan evaluasi yang lebih komprehensif karena tidak hanya memeriksa kesamaan kata secara persis, tetapi juga menilai kedekatan makna dan struktur jawaban yang dihasilkan model. Selain pengujian berbasis metrik, penelitian ini juga menerapkan User Acceptance Test (UAT) untuk mengevaluasi sistem dari perspektif pengguna. UAT dilakukan untuk menilai tingkat penerimaan, kemudahan penggunaan, serta relevansi jawaban chatbot terhadap kebutuhan layanan informasi kampus (Prasojo et al., 2024). Di sisi lain, agar memastikan proses pengujian yang terstruktur dan konsisten, penelitian ini mengimplementasikan sistem otomatisasi berbasis n8n. Platform n8n merupakan platform open-source yang memungkinkan penyusunan workflow otomatis secara terintegrasi dan dapat dijalankan berulang kali, mulai dari pengiriman pertanyaan, pengambilan respons model, hingga proses evaluasi hasil keluaran secara otomatis tanpa penulisan kode manual (McFeetors & Pant, 2022). Sistem otomatisasi ini membantu memastikan konsistensi proses pengujian sebagai landasan ilmiah dalam membandingkan performa antar model.

Berdasarkan penelitian-penelitian sebelumnya, sebagian besar chatbot layanan informasi kampus masih menggunakan pendekatan rule-based seperti AIML yang memiliki keterbatasan dalam memahami konteks pertanyaan yang kompleks dan menghasilkan jawaban yang adaptif. Selain itu, penelitian terkait implementasi Large Language Model (LLM) pada domain informasi kampus masih terbatas, khususnya dalam membandingkan performa beberapa model Generative AI pada sistem berbasis Retrieval-Augmented Generation (RAG). Penelitian sebelumnya juga belum banyak membahas pengaruh spesifikasi perangkat keras terhadap performa chatbot, terutama pada waktu respons dan efisiensi sistem. Di sisi lain, integrasi workflow automation seperti n8n untuk mendukung proses pengujian otomatis dan terstruktur masih jarang diterapkan dalam penelitian chatbot berbasis LLM. Berdasarkan research gap tersebut, kebaruan dalam penelitian ini terletak pada implementasi dan perbandingan tiga model Generative AI, yaitu GPT-OSS 20b, Qwen3-32b, dan Llama 3.1 8b pada chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG) yang terintegrasi dengan workflow automation n8n.Penelitian ini juga mengombinasikan evaluasi otomatis menggunakan Exact Match (EM) dan ROUGE dengan evaluasi pengguna melalui User Acceptance Test (UAT), serta menganalisis pengaruh spesifikasi perangkat keras terhadap performa chatbot secara keseluruhan.

Maka dari itu, penelitian ini melakukan perbandingan kinerja beberapa model Generative AI berbasis Large Language Model (LLM), yaitu GPT-OSS 20b, Qwen3-32b, dan Llama 3.1 8b dalam menjawab pertanyaan seputar dataset informasi kampus dengan menerapkan Retrieval-Augmented Generation (RAG) yang memanfaatkan vector database sebagai penyimpanan pengetahuan. Seluruh mekanisme pengujian diintegrasikan ke dalam sistem otomatisasi n8n. Dataset yang digunakan mencakup informasi penting, seperti penerimaan mahasiswa baru, biaya kuliah, lokasi kampus, dan program studi yang tersedia. Pendekatan ini diharapkan mampu mengurangi hallucination, meningkatkan relevansi jawaban, serta menghasilkan sistem layanan informasi kampus yang lebih cerdas, efisien, dan adaptif terhadap kebutuhan pengguna.

METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen komparatif dengan pendekatan kuantitatif untuk menganalisis performa beberapa model Large Language Model (LLM) pada sistem chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG). Pendekatan kuantitatif digunakan untuk mengukur tingkat performa model melalui pengujian otomatis menggunakan metrik evaluasi seperti Exact Match (EM), ROUGE Score, serta User Acceptance Testing (UAT). Analisis dilakukan secara deskriptif dengan membandingkan hasil keluaran dari masing-masing model berdasarkan akurasi jawaban, relevansi informasi, dan pengalaman pengguna dalam berinteraksi dengan sistem chatbot.

Dengan memanfaatkan pendekatan Retrieval-Augmented Generation (RAG) yang diintegrasikan dengan platform otomatisasi n8n untuk meningkatkan kemampuan model dalam menghasilkan jawaban yang relevan berdasarkan data eksternal. Metode RAG memungkinkan

model bahasa mengakses informasi tambahan dari sumber data eksternal sebelum menghasilkan respons akhir sehingga mampu meningkatkan akurasi dan relevansi jawaban (Lewis et al., 2020). Sistem RAG bekerja dengan melakukan proses pencarian informasi pada basis data vektor sebelum model menghasilkan respons akhir. Dengan pendekatan ini, model tidak hanya mengandalkan pengetahuan bawaan hasil pre-training, tetapi juga memanfaatkan informasi kontekstual dari dataset yang telah disiapkan sebelumnya sehingga mampu menghasilkan jawaban yang lebih sesuai dengan domain informasi kampus. Platform workflow automation yang digunakan dalam penelitian ini adalah n8n. N8n merupakan alat otomatisasi alur kerja (workflow automation tool) yang bersifat open-source dan dilindungi oleh lisensi fair-code yang dikembangkan sejak tahun 2019. Platform ini digunakan untuk menghubungkan berbagai layanan dan sistem cloud melalui pendekatan berbasis node sehingga mampu mengotomatisasi proses pertukaran data dan integrasi layanan secara efisien (McFeetors & Pant, 2022). Pada penelitian ini, n8n digunakan untuk mengatur alur proses chatbot mulai dari penerimaan input pengguna, proses retrieval data pada vector database, pemanggilan model LLM melalui API, hingga pengiriman respons akhir kepada pengguna secara otomatis.

Sumber data yang digunakan dalam penelitian ini berupa dataset informasi kampus yang mencakup informasi program studi, biaya pendidikan, fasilitas kampus, beasiswa, penerimaan mahasiswa baru, lokasi kampus, kegiatan akademik, dan informasi umum lainnya. Dataset diperoleh melalui teknik pengumpulan data sekunder yang berasal dari website resmi kampus, dokumen akademik, serta informasi institusi yang relevan. Data yang telah dikumpulkan kemudian diproses melalui tahapan preprocessing, chunking, dan embedding sebelum disimpan ke dalam vector database menggunakan Pinecone sebagai media penyimpanan embedding vector. Pinecone merupakan basis data vektor berbasis cloud yang dirancang untuk menyimpan, mengindeks, dan melakukan pencarian terhadap embedding vector berdimensi tinggi. Embedding tersebut merupakan representasi numerik dari data kompleks seperti teks, gambar, maupun audio. Pinecone menyediakan layanan managed vector database sehingga pengguna tidak perlu mengelola infrastruktur server dan proses scaling secara manual. Teknologi ini banyak digunakan dalam implementasi semantic search, recommendation system, Retrieval-Augmented Generation (RAG), serta penyimpanan memori jangka panjang pada sistem berbasis Artificial Intelligence (Pinecone, 2023). Pada penelitian ini, Pinecone digunakan untuk menyimpan embedding dataset informasi kampus dan melakukan similarity search guna menemukan informasi yang paling relevan terhadap pertanyaan pengguna.

Dengan menggunakan beberapa model Large Language Model (LLM) sebagai bahan perbandingan, yaitu GPT-OSS 20b, Qwen3-32b, dan LlaMA 3.1 8b. Pada Penelitian ini, Pemilihan model dilakukan berdasarkan perbedaan arsitektur, kemampuan generatif, dan efisiensi model dalam menghasilkan jawaban berbasis konteks. Arsitektur dasar LLM yang digunakan mengacu pada Transformer yang diperkenalkan oleh Ashish Vaswani. Melalui mekanisme self-attention untuk memahami hubungan antar token dalam teks (Vaswani et al., 2017). Setiap model diuji menggunakan prompt yang sama agar hasil evaluasi dapat dibandingkan secara objektif (Evidently AI, 2024). Proses pengujian dilakukan secara otomatis menggunakan workflow n8n yang mengatur alur pengiriman prompt, proses retrieval data, pemanggilan model, hingga penyimpanan hasil respons untuk dianalisis lebih lanjut.

Selain sistem chatbot berbasis AI, penelitian ini juga memanfaatkan platform deployment berbasis cloud menggunakan Vercel untuk menghosting website chatbot agar dapat diakses melalui internet. Vercel merupakan platform cloud hosting dan deployment yang mendukung berbagai framework frontend seperti React dan Next.js serta menyediakan fitur Continuous Integration dan Continuous Deployment (CI/CD) untuk mempermudah proses publikasi aplikasi website (Vercel, n.d.) Pada penelitian ini, Vercel digunakan untuk melakukan deployment website chatbot sehingga pengguna dapat berinteraksi langsung dengan sistem chatbot melalui antarmuka website secara online.

Tahapan penelitian dimulai dari pengumpulan dataset informasi kampus, preprocessing data, pembuatan embedding vector, penyimpanan data ke vector database, integrasi sistem RAG dengan workflow n8n, hingga proses pengujian model. Pada tahap preprocessing, data dibersihkan dan dipecah menjadi beberapa potongan teks (chunking) agar proses retrieval dapat berjalan lebih optimal. Selanjutnya, setiap potongan teks diubah menjadi embedding vector menggunakan model embedding dan disimpan pada vector database Pinecone. Ketika pengguna mengirimkan pertanyaan, sistem akan melakukan similarity search untuk menemukan informasi yang paling relevan, kemudian informasi tersebut diberikan sebagai konteks tambahan kepada model LLM sebelum menghasilkan jawaban akhir. Proses evaluasi dilakukan

menggunakan metode Exact Match (EM), ROUGE Score, dan User Acceptance Testing (UAT). Exact Match digunakan untuk mengukur tingkat kesamaan jawaban model dengan jawaban referensi secara tepat. EM memberikan standar evaluasi yang sangat ketat karena jawaban model harus sepenuhnya sama dengan referensi. Dalam praktiknya, proses pencocokan EM umumnya dilakukan dengan mengabaikan elemen seperti tanda baca dan kata sandang (misalnya si, sang, para, dll) agar perbandingan antara prediksi dan jawaban benar tetap adil.

Selain itu, digunakan ROUGE Score untuk mengukur tingkat kemiripan teks antara jawaban model dan jawaban referensi berdasarkan kesamaan kata atau frasa (Hamdhana, 2022). Secara prinsip, ROUGE mengevaluasi kualitas keluaran dengan mengukur jumlah unit linguistik yang tumpang tindih (overlapping units) antara teks kandidat dan referensi. Unit tersebut dapat berupa n-gram, urutan kata, maupun pasangan kata yang dipisahkan dengan jarak tertentu. Dalam varian ROUGE-N, evaluasi dilakukan berdasarkan kesesuaian n-gram antara keluaran sistem dan referensi (Holle, Munna, & Ekaputri, 2025). Sementara itu, ROUGE-L menggunakan pendekatan Longest Common Subsequence (LCS) untuk mengukur tumpang tindih berdasarkan urutan kata terpanjang yang sama antara teks kandidat dan referensi. Metode LCS tidak hanya menangkap kesamaan yang berurutan secara langsung, tetapi juga mempertimbangkan kesesuaian urutan kata yang tetap konsisten meskipun tidak selalu berdekatan. Pendekatan ini dianggap lebih fleksibel karena tidak memerlukan penentuan panjang n-gram tertentu dan mampu merepresentasikan struktur sintaktis serta koherensi kalimat dengan lebih baik (Holle et al., 2025).

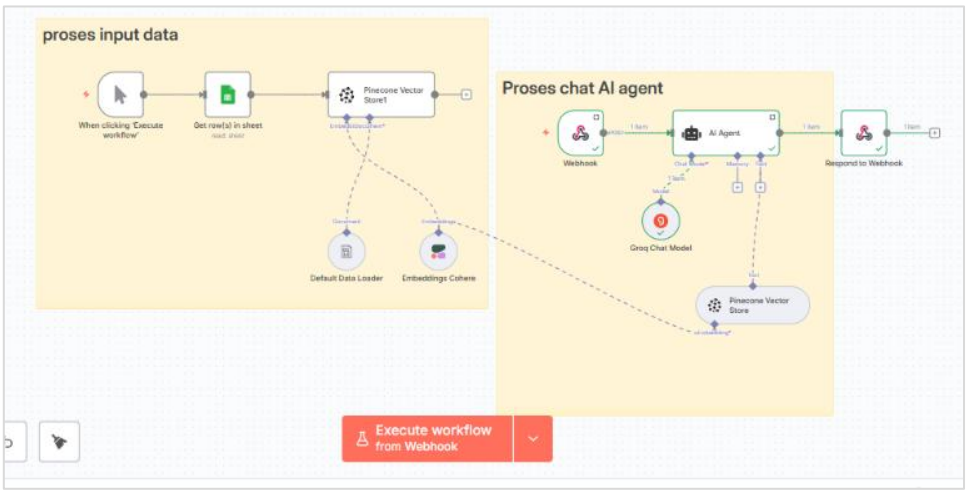
Pengujian User Acceptance Testing (UAT) dilakukan dengan melibatkan 11 mahasiswa sebagai responden yang berinteraksi langsung dengan website chatbot. Pengujian ini bertujuan untuk mengevaluasi tingkat penerimaan pengguna terhadap sistem berdasarkan aspek kemudahan penggunaan, kejelasan informasi, kecepatan respons, tampilan antarmuka, dan kepuasan pengguna secara keseluruhan. Setiap responden diminta mengisi kuesioner setelah menggunakan sistem chatbot yang telah disediakan. Metode UAT digunakan untuk memastikan bahwa sistem yang dikembangkan telah sesuai dengan kebutuhan pengguna dan dapat digunakan secara efektif dalam lingkungan nyata (Lawagita et al., 2025) Pengukuran hasil UAT menggunakan skala penilaian bertingkat yang terdiri atas lima kategori, yaitu Sangat Setuju (SS), Setuju (S), Kurang Setuju (KS), Tidak Setuju (TS), dan Tidak Jawab (TJ) dengan menggunakan penilaian skala Likert 1-5. Skala ini digunakan untuk menggambarkan tingkat kesesuaian sistem terhadap kebutuhan pengguna. Indikator pengujian UAT mencakup kesesuaian jawaban, kegunaan informasi, kemudahan pemahaman, efisiensi, serta kepuasan pengguna. Hasil penilaian selanjutnya dikonversi ke dalam bentuk persentase tingkat penerimaan untuk memberikan gambaran kuantitatif mengenai sejauh mana sistem diterima oleh pengguna (Lawagita et al., 2025).

Hasil evaluasi dari Exact Match, ROUGE Score, dan User Acceptance Testing kemudian dianalisis secara komparatif untuk menentukan model yang memiliki performa terbaik dalam menjawab pertanyaan berbasis informasi kampus. Model dengan tingkat akurasi, relevansi jawaban, dan nilai kepuasan pengguna tertinggi dianggap sebagai model yang paling optimal untuk diimplementasikan pada sistem chatbot informasi kampus berbasis Retrieval-Augmented Generation.

HASIL DAN PEMBAHASAN

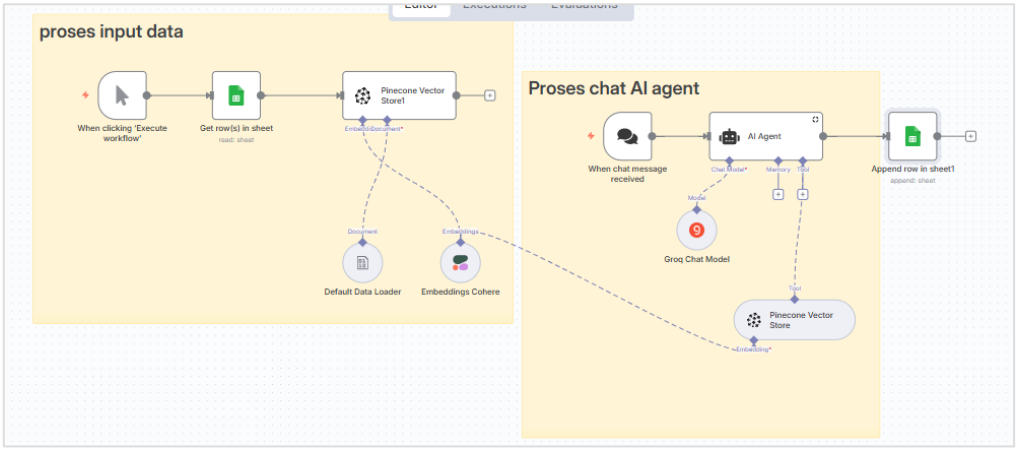
Implementasi Sistem Chatbot

Pada penelitian ini berhasil dikembangkan sistem chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG) yang terintegrasi dengan workflow automation menggunakan n8n. Sistem chatbot dirancang untuk mampu menerima pertanyaan pengguna melalui website, melakukan retrieval informasi pada vector database, kemudian menghasilkan jawaban menggunakan model Generative AI berdasarkan konteks data yang diperoleh. Implementasi workflow chatbot pada platform n8n menggunakan “node Webhook” sebagai penerima request pertanyaan dari frontend website. Node webhook menggunakan production URL sehingga dapat diakses secara publik oleh website yang telah dideploy secara online. Setelah menerima input pengguna, pertanyaan diproses oleh “node AI Agent” yang terhubung dengan model Generative AI melalui layanan Groq Chat Model. Pada tahap ini, AI Agent melakukan retrieval informasi pada Pinecone Vector Store untuk mencari data yang paling relevan berdasarkan kemiripan semantik antara pertanyaan pengguna dan dataset informasi kampus. Informasi hasil retrieval kemudian digunakan sebagai context tambahan sebelum model menghasilkan jawaban chatbot.



Gambar 1 workflow pada n8n untuk alur website chatbot

Dalam implementasinya, workflow n8n dijalankan secara lokal menggunakan Docker dan dihubungkan ke internet menggunakan layanan tunneling ngrok. Penggunaan ngrok memungkinkan sistem memperoleh URL publik HTTPS yang digunakan sebagai endpoint webhook pada website. Dengan pendekatan tersebut, website yang telah dideploy menggunakan Vercel dapat berkomunikasi dengan workflow n8n meskipun backend berjalan pada lingkungan lokal. Selain menggunakan webhook untuk integrasi website, workflow juga diuji menggunakan fitur chat internal pada n8n untuk memastikan proses retrieval dan generation berjalan dengan baik sebelum sistem diintegrasikan ke frontend website. Pengujian internal dilakukan menggunakan "node When Chat Message Received" pada n8n sehingga pengembang dapat mengevaluasi performa chatbot secara langsung pada antarmuka n8n.



Gambar 2 workflow pada n8n untuk alur chatbot dan pengujian internal

Implementasi Retrieval-Augmented Generation (RAG)

Implementasi Retrieval-Augmented Generation (RAG) dilakukan menggunakan integrasi antara Pinecone Vector Store, model embedding Cohere, dan AI Agent pada workflow n8n. Pada tahap retrieval, sistem menerima pertanyaan pengguna melalui webhook yang terhubung dengan website, kemudian pertanyaan diubah menjadi vector embedding menggunakan model "embed-multilingual-v3.0" dari Cohere. Embedding tersebut digunakan untuk melakukan similarity search terhadap dataset informasi kampus yang tersimpan pada Pinecone Vector Store. Dalam implementasinya, sistem menggunakan parameter topK sebesar 10, yang berarti sistem mengambil sepuluh data dengan tingkat kemiripan tertinggi terhadap pertanyaan pengguna. Dokumen hasil retrieval kemudian dijadikan context tambahan sebelum model Generative AI menghasilkan jawaban.

Pada workflow n8n, node Pinecone Vector Store dikonfigurasi menggunakan mode "retrieve-as-tool". Konfigurasi tersebut memungkinkan AI Agent secara otomatis menggunakan data hasil retrieval sebagai sumber informasi utama ketika menjawab pertanyaan pengguna.

Sistem juga menerapkan system prompt message untuk mengarahkan model agar menggunakan data Pinecone sebagai sumber utama jawaban, memberikan jawaban relevan sesuai konteks kampus, meminta klarifikasi apabila informasi tidak ditemukan, serta menghasilkan jawaban yang singkat dan jelas.

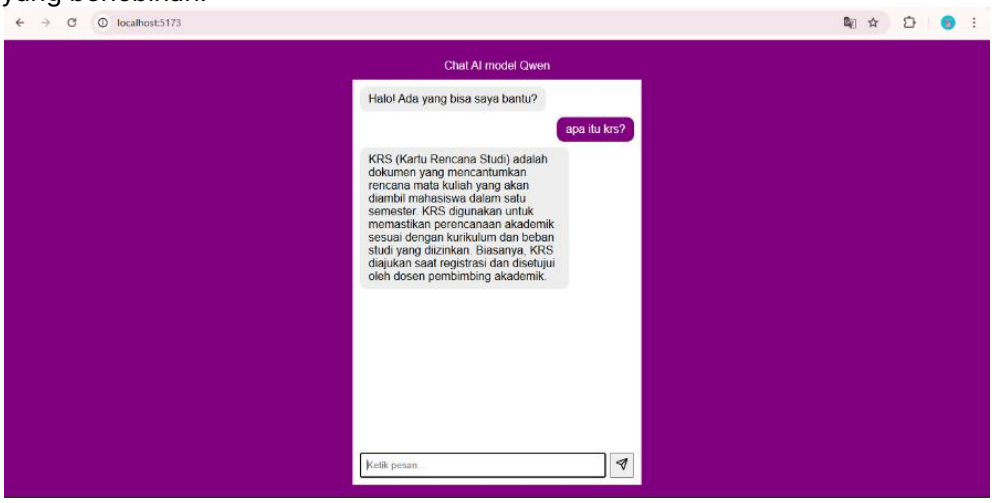
Pendekatan Retrieval-Augmented Generation membantu chatbot menghasilkan jawaban yang lebih akurat dan mengurangi kemungkinan hallucination karena model tidak hanya mengandalkan pengetahuan bawaan, tetapi juga memanfaatkan basis pengetahuan eksternal berupa dataset informasi kampus.

Implementasi Model Generative AI

Pada penelitian ini, sistem chatbot diintegrasikan dengan tiga model Generative AI, yaitu Qwen3-32b, GPT-OSS 20b, dan Llama 3.1 8b melalui layanan Groq API pada platform n8n. Integrasi model dilakukan menggunakan "node Groq Chat Model" yang terhubung langsung dengan "node AI Agent" pada workflow chatbot. Dalam implementasinya, AI Agent menerima input pertanyaan pengguna dari webhook, kemudian mengirimkan pertanyaan beserta context hasil retrieval ke model untuk diproses menjadi jawaban. Sistem juga menerapkan system prompt untuk mengontrol perilaku model agar bertindak sebagai asisten AI kampus, menggunakan data Pinecone sebagai sumber utama informasi, memberikan jawaban yang ringkas dan jelas, serta tetap fokus pada konteks informasi kampus. Selain itu, sistem membatasi panjang respons hingga maksimal 150 token agar jawaban yang dihasilkan tetap singkat dan efisien. Untuk mendukung implementasi multi-model pada proses pengujian, setiap model diimplementasikan menggunakan workflow dan path webhook yang berbeda. Pendekatan tersebut mempermudah proses integrasi model dengan website pengujian serta mempermudah evaluasi performa chatbot pada tahap User Acceptance Testing (UAT).

Implementasi Website Pengujian UAT

Implementasi website pengujian User Acceptance Test (UAT) dilakukan menggunakan library React dengan build tool Vite. Website dikembangkan sebagai antarmuka utama pengguna untuk berinteraksi secara langsung dengan chatbot informasi kampus yang telah terintegrasi dengan workflow n8n dan model Generative AI. Pada tahap deployment, website dihosting menggunakan platform Vercel sehingga dapat diakses secara online oleh responden melalui browser. Website dirancang dengan tampilan sederhana yang berfokus pada fitur chat realtime agar pengguna dapat melakukan pengujian chatbot secara langsung tanpa distraksi antarmuka yang berlebihan.



Gambar 3 tampilan website chatbot pengujian UAT

Pengujian UAT dilakukan terhadap 11 mahasiswa aktif sebagai responden. Kuesioner UAT terdiri dari 30 pertanyaan yang dibagi ke dalam lima aspek penilaian, yaitu kesesuaian jawaban, kegunaan informasi, kejelasan jawaban, efektivitas model, dan kepuasan pengguna.

Hasil Pengujian Exact Match

Pengujian Exact Match dilakukan untuk mengukur tingkat kecocokan jawaban model dengan jawaban referensi secara identik. Pada pengujian ini, nilai diberikan berdasarkan

kesamaan penuh antara jawaban yang dihasilkan model dan jawaban acuan. Hasil pengujian menunjukkan bahwa seluruh model memperoleh nilai Exact Match sebesar 0%. Hal tersebut menunjukkan bahwa tidak terdapat jawaban yang identik sepenuhnya dengan jawaban referensi karena model Generative AI cenderung menghasilkan variasi kalimat atau parafrase meskipun makna jawaban tetap relevan terhadap pertanyaan yang diberikan.

Tabel 1 hasil perhitungan exact match

Model	Skor Exact Match
Qwen3-32b	0%
GPT-OSS 20b	0%
Llama 3.1 8b	0%

Berdasarkan hasil tersebut, metrik Exact Match kurang optimal digunakan untuk mengevaluasi model Generative AI karena model cenderung menghasilkan variasi bahasa yang berbeda dengan jawaban referensi meskipun informasi yang diberikan tetap sesuai konteks.

Hasil Pengujian ROUGE

Pengujian ROUGE dilakukan untuk mengukur tingkat kesamaan jawaban model terhadap jawaban referensi berdasarkan kesamaan kata dan struktur informasi.

Tabel 2 hasil perhitungan ROUGE-1

Model	Skor Rouge-1
Qwen3-32b	69% – 70%
GPT-OSS 20b	62% – 64%
Llama 3.1 8b	69% – 72%

Berdasarkan hasil pengujian ROUGE-1, model Llama 3.1 8b memperoleh nilai tertinggi dibandingkan model lainnya. Hasil tersebut menunjukkan bahwa model mampu menghasilkan jawaban dengan tingkat kesamaan kata paling tinggi terhadap jawaban referensi. Sementara itu, model GPT-OSS 20b memperoleh nilai paling rendah karena jawaban yang dihasilkan cenderung lebih singkat dan ringkas.

Tabel 3 hasil perhitungan ROUGE-L

Model	Skor Rouge-L
Qwen3-32b	51% – 57%
GPT-OSS 20b	50% – 51%
Llama 3.1 8b	61% - 66%

Pada pengujian ROUGE-L, model Llama 3.1 8b kembali memperoleh nilai tertinggi. Hasil tersebut menunjukkan bahwa model memiliki kemampuan terbaik dalam mempertahankan struktur kalimat dan urutan informasi yang sesuai dengan jawaban referensi.

Hasil Pengujian User Acceptance Test (UAT)

Pengujian User Acceptance Test dilakukan untuk mengetahui tingkat penerimaan pengguna terhadap sistem chatbot informasi kampus yang dikembangkan. Pengujian melibatkan 11 mahasiswa aktif sebagai responden yang memberikan penilaian terhadap lima aspek, yaitu kesesuaian jawaban, kegunaan informasi, kejelasan jawaban, efektivitas model, dan kepuasan pengguna.

Tabel 4 skor penilaian UAT pada model Qwen3-32b

Aspek	Skor Persentase	Kategori
Kesesuaian jawaban	88,1%	Sangat Baik
Kegunaan informasi	89%	Sangat Baik
Kejelasan jawaban	89%	Sangat Baik
Efektivitas model	84,5%	Sangat Baik

Aspek	Skor Persentase	Kategori
Kepuasan pengguna	87,2%	Sangat Baik

Model Qwen3-32b memperoleh kategori "Sangat Baik" pada seluruh aspek penilaian. Aspek tertinggi terdapat pada kegunaan informasi dan kejelasan jawaban sebesar 89%, yang menunjukkan bahwa jawaban model dinilai jelas dan mudah dipahami pengguna.

Tabel 5 skor penilaian UAT pada model GPT-OSS 20b

Aspek	Skor Persentase	Kategori
Kesesuaian jawaban	89,9%	Sangat Baik
Kegunaan informasi	92,6%	Sangat Baik
Kejelasan jawaban	87,2%	Sangat Baik
Efektivitas model	89%	Sangat Baik
Kepuasan pengguna	89,9%	Sangat Baik

Model GPT-OSS 20b memperoleh nilai tertinggi pada aspek kegunaan informasi sebesar 92,6%. Hasil tersebut menunjukkan bahwa pengguna menilai model mampu memberikan informasi yang sangat membantu dalam memenuhi kebutuhan informasi kampus.

Tabel 6 skor penilaian UAT pada model Llama 3.1 8b

Aspek	Skor Persentase	Kategori
Kesesuaian jawaban	89%	Sangat Baik
Kegunaan informasi	86,3%	Sangat Baik
Kejelasan jawaban	87,2%	Sangat Baik
Efektivitas model	88,1%	Sangat Baik
Kepuasan pengguna	87,2%	Sangat Baik

Model Llama 3.1 8b memperoleh kategori "Sangat Baik" pada seluruh aspek penilaian. Namun, pada beberapa kondisi model mengalami error "tool_use_failed" akibat sensitivitas terhadap kompleksitas schema tool dan panjang konteks retrieval Pinecone.

Pembahasan

Berdasarkan hasil pengujian Exact Match (EM), ROUGE, waktu respons, penggunaan token, serta User Acceptance Test (UAT), setiap model Generative AI menunjukkan karakteristik performa yang berbeda dalam menghasilkan jawaban chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG).

Pada pengujian Exact Match, seluruh model memperoleh nilai 0%. Hasil tersebut menunjukkan bahwa jawaban yang dihasilkan model tidak identik secara penuh dengan jawaban referensi. Kondisi ini terjadi karena model Generative AI cenderung menghasilkan jawaban dalam bentuk parafrase atau variasi struktur kalimat meskipun informasi yang disampaikan tetap sesuai dengan konteks pertanyaan. Oleh karena itu, metrik Exact Match kurang optimal digunakan untuk mengevaluasi model chatbot berbasis Generative AI karena model tidak dirancang untuk menghasilkan jawaban yang benar-benar identik dengan expected output. Berbeda dengan Exact Match, pengujian ROUGE menunjukkan hasil evaluasi yang lebih relevan dalam mengukur kualitas jawaban model karena tidak hanya menilai kesamaan jawaban secara identik, tetapi juga mempertimbangkan kesamaan kata dan struktur informasi terhadap jawaban referensi.

Pembahasan Hasil Model Llama 3.1 8b

Berdasarkan hasil pengujian ROUGE-1 dan ROUGE-L, model Llama 3.1 8b memperoleh

performa terbaik dibandingkan model lainnya dengan nilai ROUGE-1 sebesar 69%–72% dan ROUGE-L sebesar 61%–66%. Hasil tersebut menunjukkan bahwa model mampu menghasilkan jawaban dengan tingkat kesamaan kata dan koherensi struktur informasi yang paling tinggi terhadap expected output. Tingginya nilai ROUGE, khususnya pada ROUGE-L, menunjukkan bahwa model mampu menjaga keterkaitan dan urutan informasi dalam jawaban sehingga respons yang dihasilkan lebih relevan dan terstruktur. Performa tersebut didukung oleh arsitektur transformer modern pada Llama 3.1 yang menggunakan mekanisme self-attention untuk memahami konteks secara lebih luas. Selain itu, kemampuan contextual understanding pada model membantu proses pemahaman retrieval context dari Pinecone Vector Store sehingga informasi yang diperoleh dari proses retrieval dapat dimanfaatkan secara optimal dalam pembentukan jawaban. Llama 3.1 juga telah dioptimalkan untuk instruction following sehingga mampu menghasilkan respons yang lebih sesuai dengan instruksi dan konteks pertanyaan pengguna.

Meskipun menggunakan varian model 8B dengan jumlah parameter lebih kecil dibandingkan model lainnya, Llama 3.1 8b tetap mampu menghasilkan kualitas jawaban yang baik. Hal tersebut menunjukkan bahwa ukuran parameter tidak selalu menentukan kualitas jawaban secara langsung. Efisiensi arsitektur seperti penggunaan Grouped-Query Attention (GQA) memungkinkan model tetap memiliki pemahaman konteks yang baik tanpa memerlukan jumlah parameter yang sangat besar. Selain itu, kualitas data pelatihan juga menjadi faktor penting dalam menjaga keseimbangan antara ketepatan informasi dan kealamian bahasa yang dihasilkan. Kemampuan model dalam memahami retrieval context dan menghasilkan jawaban yang relevan menjadi salah satu faktor utama dalam meningkatkan performa chatbot berbasis RAG. Namun demikian, model Llama 3.1 8b memiliki waktu respons yang relatif lebih lama, yaitu sekitar 0,26–15 detik. Kondisi ini dipengaruhi oleh kompleksitas proses inferensi pada sistem Retrieval-Augmented Generation (RAG). Sebelum menghasilkan jawaban, sistem harus melakukan pencarian semantik pada Pinecone Vector Store untuk memperoleh retrieval context yang relevan, kemudian mengintegrasikan konteks tersebut ke dalam proses generasi jawaban. Beban komputasi menjadi lebih besar ketika model harus memproses retrieval context yang panjang karena mekanisme attention pada transformer memiliki kompleksitas yang meningkat terhadap panjang input. Selain itu, proses generasi jawaban dilakukan secara autoregresif, yaitu token diproses satu per satu secara berurutan sehingga waktu inferensi menjadi lebih lambat ketika jumlah konteks yang digunakan semakin banyak.

Pada beberapa kondisi, model juga masih mengalami error “tool_use_failed” akibat sensitivitas terhadap kompleksitas schema tool, panjang retrieval context Pinecone, dan instruksi sistem yang terlalu kompleks. Model dengan parameter lebih kecil cenderung memiliki keterbatasan dalam menangani workflow yang terlalu kompleks atau format output terstruktur yang ketat dibandingkan model berukuran lebih besar. Kondisi tersebut menunjukkan bahwa meskipun Llama 3.1 8b memiliki performa yang baik dalam menghasilkan jawaban relevan, model masih memiliki keterbatasan dalam stabilitas proses inferensi ketika menghadapi retrieval context dan instruksi sistem yang kompleks.

Pembahasan Hasil Model Qwen3-32

Model Qwen3-32 menunjukkan performa yang cukup baik dengan nilai ROUGE-1 sebesar 69%–70% dan ROUGE-L sebesar 51%–57%. Berdasarkan hasil UAT, model ini memperoleh nilai tertinggi pada aspek kejelasan jawaban sebesar 89%. Hal tersebut menunjukkan bahwa jawaban yang dihasilkan model lebih mudah dipahami dan memiliki struktur bahasa yang lebih natural bagi pengguna. Selain itu, model juga mampu menghasilkan informasi yang relevan sesuai konteks pertanyaan pengguna. Karakteristik jawaban yang panjang dan deskriptif dipengaruhi oleh arsitektur Qwen3-32 yang menggunakan pendekatan transformer decoder-only sehingga model mampu mempelajari hubungan kontekstual antar token secara efektif. Selain itu, model ini dirancang menggunakan pendekatan reasoning hibrid yang menggabungkan mode respons cepat (non-thinking) dan mode penalaran mendalam (thinking). Pendekatan tersebut memungkinkan model menghasilkan jawaban yang lebih detail dan kontekstual, terutama pada pertanyaan yang membutuhkan pemahaman informasi yang lebih kompleks. Namun demikian, model Qwen3-32 memiliki waktu respons yang relatif lebih lama dibandingkan GPT-OSS 20b yaitu sekitar 0,33–11,77 detik. Waktu respons tersebut dipengaruhi oleh karakteristik model yang cenderung menghasilkan jawaban lebih panjang dan deskriptif. Struktur jawaban yang lebih panjang menyebabkan proses generasi teks membutuhkan lebih banyak token dan sumber daya komputasi selama inferensi berlangsung.

Berdasarkan hasil pengujian penggunaan token pada dashboard Groq, model Qwen3-32 menggunakan sekitar 49,7 ribu uncached input tokens dan 10,6 ribu output tokens dengan total pemrosesan sekitar 60,3 ribu token. Jumlah penggunaan token tersebut menunjukkan bahwa model membutuhkan proses komputasi yang cukup besar dalam menghasilkan jawaban chatbot.

Pembahasan Hasil Model GPT-OSS 20b

Pada model GPT-OSS 20b menunjukkan keunggulan utama pada aspek efisiensi waktu respons dan pengalaman pengguna. Model ini memperoleh waktu respons tercepat yaitu sekitar 0,31–10 detik selama pengujian berlangsung. Selain itu, model juga memperoleh nilai tertinggi pada aspek kegunaan informasi sebesar 92,6% dan kepuasan pengguna sebesar 89,9% berdasarkan hasil UAT. Hasil tersebut menunjukkan bahwa pengguna menilai model mampu memberikan informasi yang bermanfaat secara cepat dan efisien. Jawaban yang dihasilkan GPT-OSS 20b cenderung lebih singkat dan langsung menuju inti informasi sehingga pengguna dapat memperoleh jawaban dengan lebih cepat. Karakteristik tersebut menjadi salah satu faktor yang mempengaruhi tingginya tingkat kepuasan pengguna terhadap model GPT-OSS 20b. Kecepatan respons GPT-OSS 20b juga dipengaruhi oleh arsitektur Transformer yang menggunakan mekanisme self-attention sehingga memungkinkan proses pemrosesan data dilakukan secara paralel. Mekanisme tersebut membuat proses inferensi menjadi lebih cepat dan efisien dibandingkan pendekatan model sekuensial tradisional seperti Recurrent Neural Network (RNN). Selain itu, GPT-OSS 20b memanfaatkan mekanisme Key-Value Cache (KV Cache) untuk menyimpan representasi token sebelumnya di dalam memori. Dengan adanya KV Cache, model tidak perlu melakukan perhitungan ulang terhadap seluruh konteks pada setiap proses generasi token sehingga waktu komputasi dapat dikurangi secara signifikan. Kondisi tersebut turut mendukung kemampuan model dalam menghasilkan respons chatbot yang cepat meskipun memproses retrieval context dan jumlah token yang cukup besar. Namun demikian, nilai ROUGE model masih berada di bawah model Llama 3.1 8b sehingga tingkat kesamaan jawaban terhadap referensi belum sebaik model tersebut.

Berdasarkan hasil pengujian penggunaan token, GPT-OSS 20b memiliki penggunaan token paling tinggi dibandingkan model lainnya. Model ini menggunakan sekitar 60 ribu uncached input tokens dan 19,9 ribu output tokens dengan total pemrosesan sekitar 79,9 ribu token. Selain itu, terdapat sekitar 11,3 ribu cached input tokens sehingga total keseluruhan penggunaan token mencapai sekitar 91,2 ribu token. Tingginya penggunaan token menunjukkan bahwa model memproses retrieval context dalam jumlah besar dan menghasilkan output yang cukup kompleks. Kondisi tersebut dipengaruhi oleh kemampuan model dalam mempertahankan konteks percakapan dan memanfaatkan informasi retrieval secara lebih luas selama proses generation berlangsung. Selain itu, karakteristik arsitektur Transformer pada GPT-OSS 20b memungkinkan model memproses lebih banyak token secara paralel sehingga model mampu menangani konteks retrieval yang panjang tanpa mengalami penurunan performa yang signifikan. Penggunaan KV Cache juga membantu model menyimpan representasi token sebelumnya sehingga proses komputasi ulang terhadap konteks tidak perlu dilakukan pada setiap generasi token. Meskipun penggunaan token relatif tinggi, model tetap mampu memberikan waktu respons tercepat karena optimasi inferensi pada layanan Groq mampu memproses token dalam jumlah besar secara lebih cepat dan efisien.Secara keseluruhan, implementasi Retrieval-Augmented Generation (RAG) memberikan pengaruh signifikan terhadap kualitas jawaban chatbot informasi kampus. Dengan adanya retrieval context dari Pinecone Vector Store, model mampu menghasilkan jawaban yang lebih relevan, sesuai konteks, dan mengurangi kemungkinan hallucination. Sistem RAG membantu model memperoleh informasi tambahan dari dataset eksternal sebelum proses generation dilakukan sehingga chatbot tidak hanya mengandalkan pengetahuan bawaan model hasil pre-training.

Hasil penelitian menunjukkan bahwa model Llama 3.1 8b memiliki kualitas jawaban terbaik berdasarkan pengujian ROUGE, model Qwen3-32b unggul pada kejelasan jawaban, sedangkan GPT-OSS 20b memiliki performa terbaik pada aspek efisiensi waktu respons dan pengalaman pengguna. Dengan demikian, setiap model memiliki keunggulan dan kelemahan masing-masing tergantung pada kebutuhan implementasi chatbot yang diinginkan.

Faktor Yang Mempengaruhi Hasil Jawaban Model

Hasil pengujian menunjukkan bahwa kualitas dan waktu respons chatbot tidak hanya dipengaruhi oleh kemampuan model Generative AI yang digunakan, tetapi juga oleh spesifikasi

perangkat keras (hardware) yang digunakan untuk menjalankan sistem. Pada penelitian ini, proses chatbot melibatkan beberapa tahapan seperti retrieval data dari Pinecone, eksekusi workflow n8n, komunikasi API dengan model AI melalui Groq, serta generation jawaban. Seluruh proses tersebut membutuhkan sumber daya komputasi yang cukup besar sehingga spesifikasi perangkat dapat memengaruhi performa sistem secara keseluruhan.

Perbedaan spesifikasi perangkat keras terlihat memberikan pengaruh terhadap waktu respons chatbot. Pada perangkat pertama yang digunakan untuk pengujian, sistem memiliki spesifikasi berupa prosesor AMD 3020e dengan kecepatan 1.20 GHz, RAM 8 GB dengan usable memory sekitar 5,88 GB, serta GPU AMD Radeon Graphics 2 GB. Spesifikasi tersebut tergolong entry-level sehingga kemampuan pemrosesan multitasking dan komputasi masih terbatas. Kondisi ini menyebabkan proses pengolahan request, eksekusi workflow n8n, komunikasi API, serta generation jawaban model membutuhkan waktu lebih lama dengan rata-rata respons sekitar 15–23 detik.

Sebaliknya, perangkat kedua yang digunakan untuk setup workflow memiliki spesifikasi yang lebih tinggi, yaitu prosesor AMD Ryzen 7 8700G dengan 16 logical processor, RAM 64 GB, dan GPU Radeon 780M Graphics dengan kapasitas shared memory yang lebih besar. Spesifikasi tersebut memberikan performa komputasi yang lebih baik sehingga proses workflow, pengolahan request, serta komunikasi dengan model AI dapat berjalan lebih cepat dan stabil. Hasil pengujian menunjukkan bahwa waktu respons chatbot pada perangkat ini berada pada rentang sekitar 0,30–10 detik. Perbedaan waktu respons tersebut dipengaruhi oleh beberapa faktor utama. Faktor pertama adalah kemampuan prosesor (CPU). Prosesor dengan jumlah core dan thread yang lebih banyak mampu menangani eksekusi workflow, request API, serta proses paralel dengan lebih cepat. AMD Ryzen 7 8700G memiliki performa yang jauh lebih tinggi dibandingkan AMD 3020e sehingga latensi pemrosesan sistem menjadi lebih rendah.

Faktor kedua adalah kapasitas RAM yang digunakan. RAM dengan kapasitas besar memungkinkan sistem menjalankan Docker, n8n, browser, serta proses chatbot secara bersamaan tanpa mengalami bottleneck memori. Pada perangkat dengan RAM 8 GB, penggunaan memori yang hampir penuh menyebabkan performa sistem menurun sehingga waktu respons chatbot menjadi lebih lambat. Faktor berikutnya adalah kemampuan GPU terintegrasi. GPU Radeon 780M Graphics pada perangkat kedua memiliki performa grafis dan komputasi yang lebih baik dibandingkan Radeon Graphics pada AMD 3020e. Kondisi tersebut membantu stabilitas sistem dan mempercepat proses komputasi saat workflow chatbot berjalan. Selain itu, kecepatan sistem secara keseluruhan juga memengaruhi performa chatbot. Perangkat dengan spesifikasi tinggi memiliki kemampuan membaca data, menjalankan container Docker, dan memproses jaringan dengan lebih cepat sehingga latensi end-to-end chatbot menjadi lebih kecil.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa spesifikasi perangkat keras memiliki pengaruh signifikan terhadap performa chatbot, khususnya pada waktu respons sistem. Semakin tinggi kemampuan CPU, RAM, dan GPU yang digunakan, maka semakin cepat pula proses retrieval dan generation jawaban yang dihasilkan chatbot.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, implementasi chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG) menggunakan workflow automation n8n berhasil dikembangkan dan mampu menghasilkan jawaban berbasis konteks informasi kampus secara otomatis. Implementasi RAG membantu model memperoleh retrieval context dari Pinecone Vector Store sebelum proses generation dilakukan sehingga mampu meningkatkan relevansi jawaban dan mengurangi kemungkinan hallucination pada chatbot. Hasil pengujian menunjukkan bahwa seluruh model Generative AI menghasilkan nilai Exact Match sebesar 0% karena model cenderung menghasilkan variasi jawaban berbentuk parafrase meskipun informasi yang diberikan tetap relevan dengan jawaban referensi. Pada pengujian ROUGE, model Llama 3.1 8b memperoleh performa terbaik dengan nilai ROUGE-1 sebesar 69%–72% dan ROUGE-L sebesar 61%–66%, yang menunjukkan tingkat kesamaan jawaban tertinggi terhadap jawaban referensi. Sementara itu, model GPT-OSS 20b memperoleh tingkat kepuasan pengguna tertinggi serta waktu respons tercepat berdasarkan hasil User Acceptance Test (UAT), sedangkan model Qwen3-32b menunjukkan keunggulan pada aspek kejelasan jawaban.

Perbedaan hasil performa antar model dipengaruhi oleh beberapa faktor, seperti kemampuan model dalam memahami retrieval context, struktur jawaban yang dihasilkan, kompleksitas proses inferensi, jumlah penggunaan token, serta sensitivitas model terhadap

panjang konteks dan schema tool pada workflow n8n. Selain itu, hasil penelitian juga menunjukkan bahwa spesifikasi perangkat keras (hardware) memiliki pengaruh signifikan terhadap performa chatbot, khususnya pada waktu respons dan efisiensi proses generation jawaban. Perangkat dengan spesifikasi lebih tinggi, seperti prosesor multi-core, kapasitas RAM besar, dan GPU yang lebih baik mampu menjalankan workflow n8n, retrieval Pinecone, komunikasi API, serta proses generation model dengan lebih cepat dan stabil dibandingkan perangkat dengan spesifikasi rendah. Kondisi tersebut menyebabkan sistem memiliki latensi yang lebih rendah sehingga chatbot mampu memberikan respons lebih cepat dan efisien.

Berdasarkan hasil penelitian, setiap model memiliki karakteristik performa yang berbeda sehingga pemilihan model chatbot dapat disesuaikan dengan kebutuhan implementasi sistem. Model Llama 3.1 8b lebih unggul pada kualitas dan relevansi jawaban, GPT-OSS 20b lebih optimal pada efisiensi dan pengalaman pengguna, sedangkan Qwen3-32b lebih baik dalam menghasilkan jawaban yang jelas dan mudah dipahami. Penelitian selanjutnya dapat dilakukan dengan menambahkan dataset yang lebih besar, mengoptimalkan proses retrieval, meningkatkan kualitas perangkat keras, serta melakukan fine-tuning model untuk meningkatkan performa chatbot informasi kampus berbasis Generative AI.

UCAPAN TERIMA KASIH

Penulis mengucapkan puji dan syukur kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Penulis juga mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, masukan, serta bimbingan selama proses penelitian dan penyusunan artikel ini. Ucapan terima kasih turut disampaikan kepada kedua orang tua dan keluarga yang selalu memberikan doa, dukungan, motivasi, serta semangat kepada penulis selama proses penelitian berlangsung sehingga penulis dapat menyelesaikan penelitian ini dengan baik. Selain itu, penulis juga mengucapkan terima kasih kepada seluruh responden mahasiswa yang telah bersedia mengikuti proses User Acceptance Test (UAT) sehingga penelitian dapat berjalan dengan lancar.

Penulis turut menyampaikan apresiasi kepada pihak kampus yang telah menyediakan dukungan fasilitas serta sumber data yang diperlukan dalam pengembangan sistem chatbot informasi kampus berbasis Retrieval-Augmented Generation (RAG). Semoga penelitian ini dapat memberikan manfaat dan kontribusi dalam pengembangan teknologi chatbot berbasis Generative Artificial Intelligence di bidang pendidikan.

REFERENSI

Aliphadji Talaohu, S., Soekarta, R., & Surahmanto, M. (2025). Implementasi LLM Pada Chatbot PMB Universitas Muhammadiyah Sorong Menggunakan Metode RAG Berbasis Website, *03*(02).

Attigeri, G., Agrawal, A., & Kolekar, S. V. (2024). Advanced NLP Models for Technical University Information Chatbots: Development and Comparative Analysis. *IEEE Access*, *12*, 29633–29647. doi:10.1109/ACCESS.2024.3368382

Evidently AI. (2024). LLM Evaluation Metrics. Retrieved from https://www.evidentlyai.com/llm-guide/llm-evaluation-metrics#ranking-metrics

Hamdhana, D. (2022). Memahami ROUGE sebagai Evaluation Metric untuk NLP. Retrieved from https://defryhamdhana.medium.com/memahami-rouge-sebagai-evaluation-matric-untuk-nlp-a08605878da4

Holle, K. F. H., Munna, D. N., & Ekaputri, E. W. (2025). Performance Evaluation of Transformer Models: Scratch, Bart, and Bert for News Document Summarization. *Jurnal Teknik Informatika (Jutif)*, *6*(2), 787–802. doi:10.52436/1.jutif.2025.6.2.2534

Lawagita, D., Putri, A., Afriansyah, R., Prayesy, P. A., Manufaktur, P., Bangka, N., … Sungailiat Bangka, K. (2025). *Implementasi User Acceptance Testing (UAT) Pada Pengujian Sistem Informasi Akademik dan Keuangan Santri*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Retrieved from https://github.com/huggingface/transformers/blob/master/

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., … Zhao, L. (2025). Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *ACM Computing Surveys*. doi:10.1145/3764579

McFeetors, Jason., & Pant, Tanay. (2022). *Rapid Product Development with N8n : Practical Guide to Creating Digital Products on the Web Using Workflow Automation and N8n*. Packt Publishing, Limited.

Pinecone. (2023). What is a Vector Database & How Does it Work? Use Cases + Examples. Retrieved 10 December 2025, from https://www.pinecone.io/learn/vector-database/

Prasojo, B., Huda, M., Khasanah, I. N., & Wahyuningsih, E. (2024). APLIKASI CHATBOT BERBASIS TELEGRAM UNTUK LAYANAN INFORMASI DAN AKADEMIK KAMPUS UNIVERSITAS MA'ARIF NAHDLATUL ULAMA KEBUMEN. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2). doi:10.23960/jitet.v12i2.4013

Risch, J., Möller, T., Gutsch, J., & Pietsch, M. (2021). *Semantic Answer Similarity for Evaluating Question Answering Models*. Retrieved from https://semantic-answer-similarity.s3.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., … Polosukhin, I. (2017). *Attention Is All You Need*.

Vercel. (n.d.). Vercel. Retrieved 14 May 2026, from https://vercel.com

Zubaedah, R. and H. R. and N. M. and W. R. R. K. K. and S. I. K. D. G. and H. H. and V. T. and W. A. and I. T. and S. S. and others. (2026). *Generative Artificial Intelligence*. PT. Sonpedia Publishing Indonesia.