Penerapan RAG pada Chatbot Edukasi dan Deteksi Dini Penyakit THT

Khairina Eka Setyaputri [a,1,*], David Fitrianto [b,2], Wicaksono Yuli Sulistyo [b,3]

[a,b,c]PJJ Sistem Informasi, Universitas Siber Muhammadiyah, Pakuncen, Wirobrajan, Yogyakarta, Daerah Istimewa Yogyakarta 55253
[1] khairinaekasetyaputri@sibermu.ac.id *; [2] david2022020005@sibermu.ac.id; [3] wicaksono@sibermu.ac.id
* Korespondensi penulis

Submission:18/12/2025, Revision: 10/03/2026, Accepted : 26/04/2026

Abstract

Limited access to fast and accurate medical information is often a major constraint in the early detection of Ear, Nose, and Throat (ENT) diseases. This study proposes the development of an intelligent chatbot using the Retrieval-Augmented Generation (RAG) architecture to minimize hallucinations in Large Language Models (LLMs). The system is built using the n8n low-code automation platform, integrated with the WhatsApp API as the user interface and the gpt-4o-mini model as the inference engine. The system's knowledge base is sourced from external databases, including clinical references and visit data, processed through a vector store to ensure that responses remain within the context of valid data. Testing results indicate that the implementation of RAG increases information accuracy compared to the standard model. Furthermore, the use of n8n has proven to provide operational cost efficiency and accelerate the deployment cycle. The system successfully achieved an average latency of under 5 seconds with a Success Rate of 95%. This study concludes that the integration of RAG on a no-code platform is an effective solution for providing precise and economical health informatics services.

Keywords: Chatbot, RAG, n8n, WhatsApp

Abstrak

Keterbatasan akses informasi medis yang akurat dan cepat sering kali menjadi kendala dalam deteksi dini penyakit THT (Telinga, Hidung, dan Tenggorokan). Penelitian ini mengusulkan pengembangan chatbot cerdas menggunakan arsitektur *Retrieval-Augmented Generation* (RAG) untuk meminimalisir halusinasi pada *Large Language Model* (LLM). Sistem dibangun menggunakan platform otomasi *low-code* n8n yang diintegrasikan dengan API WhatsApp sebagai antarmuka pengguna dan model gpt-4o-mini sebagai mesin inferensi. Pengetahuan sistem bersumber dari basis data eksternal berupa data kunjungan dan referensi klinis yang diproses melalui *vector store* untuk memastikan jawaban tetap berada dalam konteks data yang sahih. Hasil pengujian menunjukkan bahwa implementasi RAG meningkatkan akurasi informasi dibandingkan model standar. Penggunaan n8n terbukti memberikan efisiensi biaya operasional dan mempercepat siklus pengembangan (*deployment*). Sistem ini berhasil mencapai rata-rata latensi di bawah 5 detik dengan tingkat keberhasilan respons (*Success Rate*) mencapai 95%. Penelitian ini menyimpulkan bahwa integrasi RAG pada platform *no-code* merupakan solusi efektif untuk penyediaan layanan informatika kesehatan yang presisi dan ekonomis.

Kata kunci: Chatbot, RAG, n8n, WhatsApp

1. Pendahuluan

Transformasi digital dalam sektor layanan kesehatan kini tidak lagi hanya berfokus pada digitalisasi rekam medis, tetapi telah berkembang ke arah pemanfaatan teknologi yang lebih interaktif dan adaptif. Salah satu perkembangan yang cukup menonjol adalah hadirnya asisten virtual berbasis kecerdasan buatan yang

dapat membantu proses triase awal sekaligus memberikan edukasi kesehatan secara mandiri kepada pasien [1]. Melalui sistem ini, pengguna dapat menyampaikan keluhan atau gejala yang dirasakan, kemudian sistem akan merespons dengan informasi awal yang sesuai. Hadirnya teknologi ini menjadi alternatif dalam mendukung layanan kesehatan, terutama ketika akses terhadap tenaga medis terbatas, sekaligus membantu masyarakat memperoleh informasi kesehatan dengan lebih mudah dan cepat [2].

Pada instalasi spesialis seperti Klinik THT (Telinga, Hidung, dan Tenggorokan) efisiensi dalam penyampaian informasi menjadi hal yang krusial, terutama seiring dengan tingginya jumlah kunjungan pasien [3]. Berdasarkan data operasional RS Mitra Sehat Medika, Pandaan, volume kunjungan menunjukkan tren yang signifikan dengan sebanyak 504 pasien tercatat pada tahun 2021 dan 593 pasien pada tahun 2022 [4]. Tren kenaikan ini menunjukkan bahwa kebutuhan layanan tidak hanya terbatas pada pemeriksaan dan pengobatan, tetapi juga mencakup kebutuhan akan informasi yang cepat dan mudah dipahami oleh pasien [5]. Proses edukasi yang masih dilakukan secara manual berpotensi memakan waktu dan kurang efisien, terutama ketika harus melayani banyak pasien dalam waktu yang bersamaan. Oleh karena itu, diperlukan suatu pendekatan yang dapat membantu menyampaikan informasi dasar secara lebih sistematis dan konsisten, sehingga pasien tetap mendapatkan pemahaman yang memadai untuk melakukan deteksi dini mengenai sakit yang dialami [6].

Seiring dengan semakin pesatnya perkembangan tekonologi di era digital, pengembangan agen percakapan (chatbot) di bidang kesehatan juga mengalami perubahan yang cukup signifikan. Jika sebelumnya chatbot banyak dibangun menggunakan pendekatan berbasis aturan (rule-based) yang cenderung kaku dan terbatas pada skenario tertentu, kini teknologi tersebut telah berkembang dengan memanfaatkan kecerdasan buatan generatif [7]. Pendekatan ini memungkinkan sistem untuk memahami konteks percakapan secara lebih fleksibel dan menghasilkan respons yang lebih natural, sehingga interaksi dengan pengguna menjadi lebih intuitif. Meskipun demikian penggunaan Large Language Models (LLM) seperti GPT-4 masih memiliki sejumlah tantangan, khususnya dalam konteks kesehatan. Salah satu permasalahan utama adalah potensi terjadinya "halusinasi" medis, yaitu kondisi ketika model menghasilkan informasi yang terdengar meyakinkan tetapi tidak sepenuhnya akurat [8]. Model juga memiliki keterbatasan dalam mengakses data lokal atau informasi spesifik yang bersifat dinamis, seperti gejala dan cara pengobatan awal suatu penyakit tertentu atau data terkini yang terus diperbarui. Hal ini menjadi perhatian penting karena informasi yang tidak tepat dapat berdampak pada pemahaman dan pengambilan keputusan oleh pengguna. Untuk mengatasi keterbatasan tersebut, pendekatan Retrieval-Augmented Generation (RAG) yang kini mulai banyak digunakan sebagai solusi[9]. Arsitektur ini mengombinasikan kemampuan generatif dari model bahasa dengan mekanisme pencarian informasi dari sumber eksternal yang relevan. Dengan demikian, sebelum menghasilkan jawaban, sistem terlebih dahulu mengambil referensi dari dokumen yang telah disediakan, sehingga respons yang diberikan menjadi lebih terarah, dapat dipertanggungjawabkan dan sesuai dengan konteks kebutuhan pengguna [10]. Pendekatan ini dinilai mampu meningkatkan kualitas dan keandalan chatbot, terutama bidang kesehatan yang menuntut tingkat akurasi yang tinggi.

Namun demikian, hasil identifikasi terhadap celah penelitian menunjukkan bahwa sebagian besar implementasi Retrieval-Augmented Generation (RAG) di sektor medis saat ini masih memiliki sejumlah keterbatasan dari sisi praktis. Banyak solusi yang dikembangkan sangat bergantung pada infrastruktur dengan tingkat kompleksitas tinggi, mulai dari kebutuhan pengembangan berbasis high-coding, penggunaan *vector database* yang relatif mahal, hingga ketergantungan pada tim teknis khusus untuk proses pengelolaan dan pemeliharaannya [11][12]. Kondisi ini menjadikan implementasi teknologi tersebut tidak mudah diadopsi oleh seluruh institusi kesehatan, khususnya bagi fasilitas dengan skala menengah ke bawah yang memiliki keterbatasan sumber daya. Proses penggabungan berbagai komponen teknologi, seperti model bahasa, basis data, serta antarmuka pengguna, sering kali membutuhkan waktu dan keahlian teknis yang tidak sederhana. Hal ini berpotensi menghambat proses implementasi dan memperlambat pemanfaatan teknologi secara optimal di lingkungan layanan kesehatan.

Penelitian ini hadir untuk mengisi celah tersebut dengan mengusulkan suatu pendekatan baru berupa arsitektur Retrieval-Augmented Generation (RAG) yang dibangun di atas platform otomasi *no-code* n8n. Pendekatan ini dirancang untuk menyederhanakan proses pengembangan sistem cerdas tanpa mengurangi kualitas fungsionalitas yang dihasilkan. Dengan memanfaatkan model *gpt-4o-mini* yang relatif efisien dari sisi biaya komputasi, sistem yang dikembangkan mampu memberikan respons yang tetap relevan dan kontekstual, namun dengan kebutuhan sumber daya yang lebih ringan dibandingkan model berskala besar lainnya. Integrasi sistem ini juga dilakukan secara langsung ke dalam ekosistem WhatsApp, yang dikenal sebagai platform komunikasi dengan tingkat pengguna yang sangat tinggi di Indonesia [13], sehingga memudahkan akses bagi masyarakat tanpa perlu menggunakan aplikasi tambahan. Penggunaan n8n sebagai platform otomasi memungkinkan proses integrasi basis pengetahuan medis dilakukan secara lebih cepat dan terstruktur tanpa memerlukan penulisan kode yang kompleks. Alur kerja yang dibangun dapat dikonfigurasi melalui antarmuka visual, sehingga mempermudah proses pengembangan, pengujian, hingga pemeliharaan sistem[14]. Dengan

pendekatan tersebut, sistem yang diusulkan tidak hanya berfokus pada aspek kecerdasan dalam menghasilkan respons, tetapi juga pada aspek kemudahan implementasi dan skalabilitas. Institusi kesehatan dengan sumber daya terbatas tetap memiliki peluang untuk mengadopsi teknologi ini secara lebih praktis. Selain itu, pemanfaatan platform yang sudah akrab di masyarakat diharapkan dapat meningkatkan tingkat penerimaan dan penggunaan sistem, sehingga solusi yang dikembangkan menjadi lebih aplikatif dan berdampak nyata dalam mendukung layanan kesehatan.

Tujuan utama dari penelitian ini adalah untuk merancang, mengimplementasikan, serta mengevaluasi kinerja sistem chatbot berbasis Retrieval-Augmented Generation (RAG) dalam konteks edukasi dan deteksi dini penyakit THT. Perancangan sistem difokuskan pada bagaimana mengintegrasikan model bahasa dengan sumber pengetahuan medis yang relevan sehingga mampu menghasilkan respons yang informatif dan sesuai dengan kebutuhan pengguna. Tahap implementasi dilakukan dengan memanfaatkan pendekatan yang efisien dan mudah diterapkan, sehingga sistem tidak hanya berfungsi secara optimal tetapi juga dapat diadopsi oleh institusi dengan keterbatasan sumber daya.

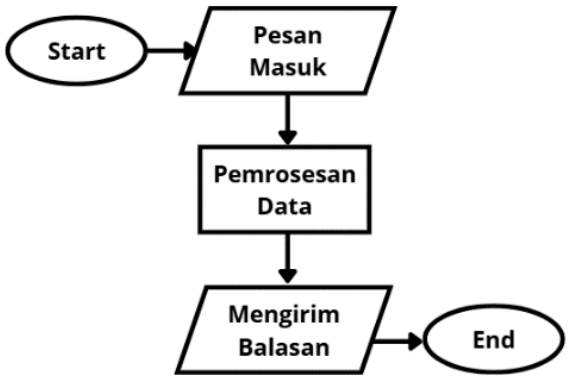
Selanjutnya, proses evaluasi dilakukan dengan memfokuskan pada performa teknis maupun aspek penerimaan pengguna. Dari sisi teknis, pengujian mencakup efisiensi waktu respons (latensi) untuk memastikan sistem dapat memberikan jawaban secara cepat, serta tingkat keberhasilan jawaban terhadap dokumen referensi (*success rate*) guna menilai sejauh mana akurasi dan relevansi informasi yang dihasilkan. Di samping itu, evaluasi juga dilakukan melalui metode *User Acceptance Testing* (UAT) untuk mengetahui tingkat kemudahan penggunaan, kejelasan informasi, serta kepuasan pengguna terhadap sistem yang dikembangkan [15]. Melalui rangkaian tahapan tersebut, penelitian ini diharapkan dapat menghasilkan sebuah model prototipe sistem chatbot cerdas yang tidak hanya unggul dari sisi akurasi informasi, tetapi juga memiliki keunggulan dalam hal efisiensi biaya, kemudahan implementasi, dan fleksibilitas pengembangan. Dengan demikian, sistem yang diusulkan berpotensi untuk direplikasi dan diterapkan pada berbagai unit layanan kesehatan lainnya, khususnya dalam mendukung penyediaan informasi kesehatan yang lebih merata dan mudah diakses oleh masyarakat.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental dengan tujuan merancang dan menguji sistem chatbot berbasis Retrieval-Augmented Generation (RAG) yang dibangun menggunakan platform n8n dan terintegrasi dengan WhatsApp. Metode yang digunakan berfokus pada implementasi sistem secara langsung, mulai dari pengolahan data, integrasi komponen, hingga pengujian performa

2.1. Desain Sistem

Sistem dikembangkan menggunakan n8n sebagai *workflow orchestration engine* yang mengatur alur proses komunikasi antara pengguna, basis pengetahuan, dan model bahasa. Integrasi dengan WhatsApp dilakukan melalui API sebagai media input dan output sistem. Model bahasa yang digunakan adalah *gpt-4o-mini*, yang dipilih berdasarkan pertimbangan efisiensi biaya dan kecepatan respons. Adapun *flowchart* dari *chatbot* yang akan dikembangkan dapat dilihat pada Gambar 1 sebagai berikut.



Gambar 1. *Flowchart*

Flowchart akan memproses pesan masuk yang dikirimkan oleh pengguna, memproses pesan dengan *AI Agent* dan mengirimkan balasan yang sesuai dengan data yang dimiliki oleh sistem kepada pengguna.

2.2. Mekanisme Retrieval-Augmented Generation (RAG)

Proses RAG pada sistem ini terdiri dari dua tahap utama, yaitu proses pencarian dokumen relevan (*retrieval*) dan proses generasi jawaban (*generation*). Tahap retrieval dilakukan dengan menghitung tingkat kemiripan antara pertanyaan pengguna dan data dalam basis pengetahuan menggunakan pendekatan *cosine similarity* berbasis representasi vektor teks.
Nilai kemiripan antara dua vektor dihitung menggunakan persamaan berikut:

$$\text{Similarity } (A, B) = \frac{\text{A}\cdot\text{B}}{||\text{A}||\ ||B||}$$

Pada persamaan tersebut, A merupakan vektor representasi dari pertanyaan pengguna, sedangkan B merupakan vektor dari dokumen dalam basis pengetahuan. Notasi A·B menyatakan hasil perkalian dot product antara kedua vektor, dan ‖A‖ serta ‖B‖ masing-masing menyatakan panjang (norma) dari vektor tersebut. Nilai similarity digunakan untuk menentukan dokumen yang paling relevan terhadap pertanyaan pengguna.

Dokumen dengan nilai kemiripan tertinggi kemudian digunakan sebagai konteks tambahan yang dikirimkan ke model LLM pada tahap generasi. Dengan demikian, jawaban yang dihasilkan tidak hanya bergantung pada kemampuan generatif model, tetapi juga didasarkan pada data referensi yang tersedia.

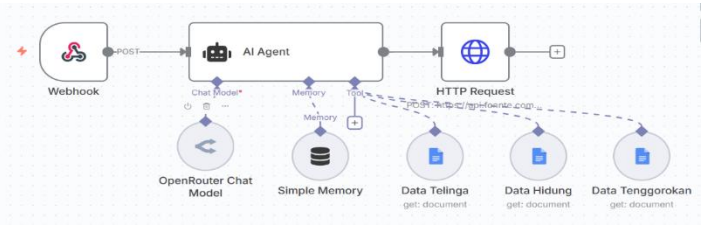
2.3. Pengujian

Pengujian dilakukan menggunakan 50 variasi pertanyaan yang mewakili kondisi umum pada layanan THT. Setiap pertanyaan diuji sebanyak tiga kali untuk melihat konsistensi hasil, sehingga total pengujian mencapai 150 percobaan meliputi waktu respons (latensi), tingkat keberhasilan (success rate), tingkat penerimaan pengguna (UAT).

3. Hasil dan Pembahasan

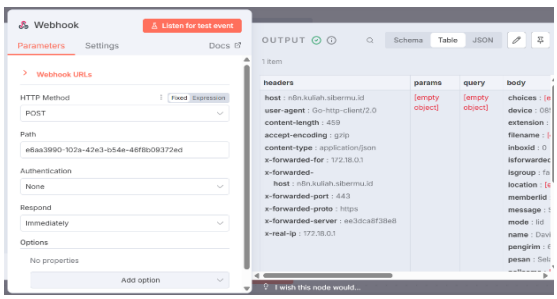
3.1. Implementasi Sistem pada n8n dan WhatsApp

Implementasi sistem dilakukan dengan memanfaatkan n8n sebagai *workflow engine* utama yang mengatur seluruh proses komunikasi, pengolahan data, serta integrasi dengan layanan eksternal. Sistem dibangun dalam bentuk satu *workflow* utama yang terdiri dari beberapa node yang saling terhubung dan berjalan secara berurutan. Adapun alur kerja pada *workflow* n8n dapat dilihat pada Gambar 2.



Gambar 2. *Workflow* n8n

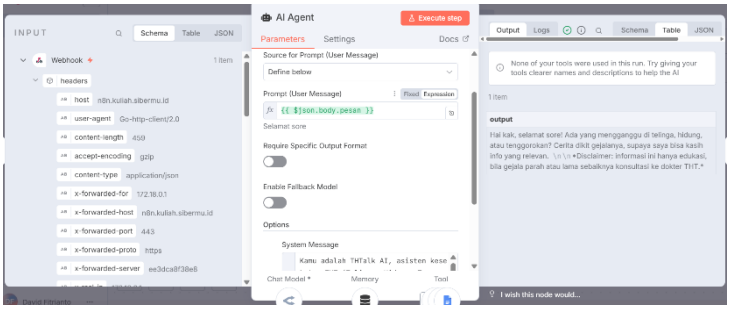
Alur dimulai dari Webhook Node yang berperan sebagai titik masuk sistem. Node ini menerima permintaan HTTP POST dari layanan WhatsApp API yang berisi pesan pengguna.



Gambar 3. Konfigurasi Webhook

Pada Gambar 3, menggambarkan konfigurasi pada whebhook sebagai langkah awal penerima pesan ke dalam sistem n8n. Payload yang diterima mencakup isi pesan, nomor pengirim, serta metadata lainnya. Data ini kemudian diteruskan ke node berikutnya tanpa proses manual tambahan.

Pesan yang masuk kemudian diproses oleh AI Agent Node, yang menjadi inti dari sistem. Node ini dikonfigurasi untuk mengelola interaksi dengan model bahasa sekaligus mengatur penggunaan *tools* dan *memory*.



Gambar 4. Konfirgurasi AI Agent

Berdasarkan pada Gambar 4, pada bagian AI Agent ini terdapat tiga konfigurasi utama:

1. Chat Model (OpenRouter Chat Model)
   Digunakan sebagai penghubung ke model *gpt-4o-mini*. Model ini bertugas menghasilkan respons berdasarkan input dan konteks yang diberikan.
2. Memory (Simple Memory)
   Digunakan untuk menyimpan riwayat percakapan dalam satu sesi. Dengan adanya komponen ini, sistem mampu mempertahankan konteks dialog sehingga respons menjadi lebih konsisten terhadap percakapan sebelumnya.
3. Tools (Data Retrieval)
   Terdiri dari tiga sumber data dalam bentuk google docs, yaitu:
   a) Data Telinga
   b) Data Hidung
   c) Data Tenggorokan

Ketiga *tools* ini berfungsi sebagai sumber pengetahuan eksternal yang diakses saat proses RAG berlangsung. Masing-masing tool dikonfigurasi untuk mengambil dokumen dari dataset yang relevan.

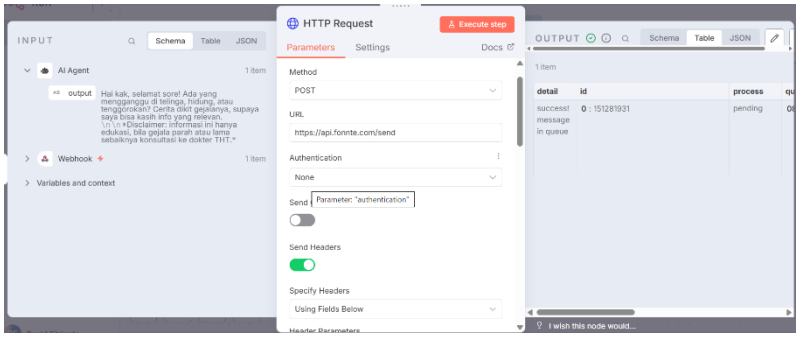
Berbeda dengan implementasi RAG berbasis *vector database*, sistem ini menggunakan pendekatan berbasis pemanggilan dokumen langsung melalui *tools* di dalam AI Agent. Ketika pengguna mengajukan pertanyaan, AI Agent akan:

1. Menganalisis intent dari pertanyaan.
2. Menentukan sumber data yang relevan (telinga, hidung, atau tenggorokan).
3. Mengakses dokumen melalui tool yang sesuai.
4. Menggunakan hasil retrieval sebagai konteks tambahan dalam proses generasi jawaban.

Pendekatan ini memungkinkan sistem tetap menjalankan prinsip RAG tanpa memerlukan infrastruktur kompleks seperti embedding pipeline atau vector store.

Setelah konteks diperoleh, AI Agent mengirimkan permintaan ke model bahasa melalui koneksi OpenRouter. Model kemudian menghasilkan respons berdasarkan kombinasi antara:

1. Pertanyaan pengguna
2. Konteks hasil retrieval
3. Riwayat percakapan (memory)



Gambar 5. Konfigurasi HTTP Request

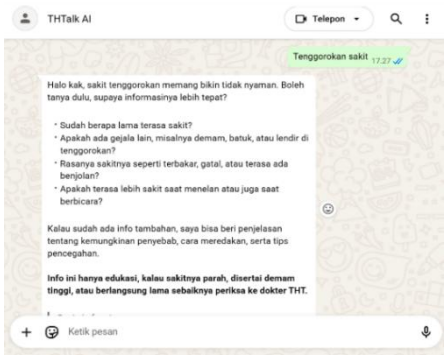
Gambar 5. merupakan gambaran konfigurasi pada HTTP Request yang telah dikonfigurasi untuk mengirim pesan kembali ke pengguna melalui WhatsApp API (dalam hal ini menggunakan endpoint seperti Fonnte API). Hasil respons dari model diteruskan ke HTTP Request Node sebagai output dari proses yang telah dikerjakan dan dikirimkan kepada pengguna sebagai jawaban dari pertanyaan yang diberikan. Node ini melakukan:

1. Formatting pesan ke dalam struktur JSON
2. Penyesuaian nomor tujuan
3. Pengiriman pesan secara real-time

Dengan demikian, seluruh proses dari input hingga output berjalan dalam satu alur otomatis tanpa intervensi manual.

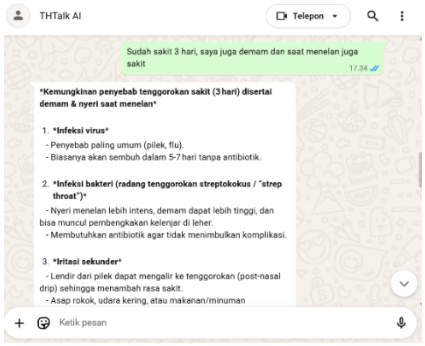
Gambar 6. Uji Coba Selamat Sore

Berdasarkan pada gambar 5, sistem yang telah dibangun dengan n8n berhasil menjawab pesan masuk melalui WhatsApp secara otomatis, Pada saat pertama kali pesan masuk chatbot akan langsung memberikan pesan selamat datang dan disclaimer.



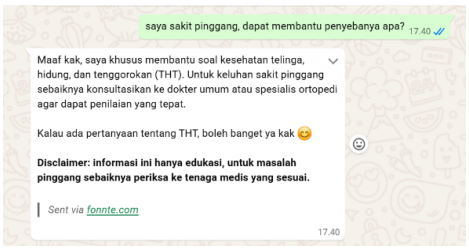
Gambar 6. Pertanyaan Umum

Gambar 6, chatbot diuji coba dengan mengirimkan pertanyaan yang umum tentang THT. Chatbot dapat secara otomatis menjawab pertanyaan dengan memberikan pertanyaan balik yang dimaksudkan untuk mencari informasi mengenai sakit atau keluhan yang dialami.



Gambar 7. Informasi Detail

Ketika chatbot telah mendapatkan informasi lebih detail tentang sakit yang dialami seperti digambarkan pada Gambar 7, chatbot akan menjawab sesuai dengan informasi yang ada pada data yang telah tersimpan di tols n8n. Informasi ini disesuaikan dengan topik yang ditanyakan oleh pegguna yang meliputi telinga, hidung atau tenggorokan.



Gambar 8. Pertanyaan Luar Topik THT

Pertanyaan diluar topik THT seperti pada gambar 8, chatbot tidak akan memberikan informasi yang kompleks karena adanya pembatasan yang diterapkan di AI Agent. Chatbot akan memberikan saran kepada pengguna untuk berkonsultasi dengan dokter atau fasilitas kesehatan sesuai dengan topik. Pada setiap akhir pesan jawaban yang dikirimkan kepada pengguna chatbot akan memberikan disclaimer bahwasannya informasi yang diberikan oleh sistem ini hanya sebagai edukasi dan alat deteksi dini. Untuk konsultasi dan penanganan lebih lanjut chatbot tetap menyarankan untuk mengunjungi fasilitas kesehatan terdekat.

3.2. Pengujian Chatbot

Pengujian dilakukan untuk melihat sejauh mana sistem mampu bekerja ketika menerima dan memproses permintaan dari pengguna. Penilaian tidak hanya dilihat dari jawaban yang dihasilkan, tetapi juga dari proses yang terjadi di dalam sistem, mulai dari penerimaan input, pencarian data, hingga pengiriman respons kembali ke pengguna. Setiap skenario diuji lebih dari satu kali untuk memastikan hasil yang diberikan tidak berubah secara signifikan. Selama proses ini, sistem diamati dari beberapa sisi, seperti kecepatan dalam memberikan respons, kesesuaian jawaban dengan data yang digunakan sebagai referensi, serta kelancaran alur kerja pada n8n ketika menjalankan setiap proses. Adapun rekap hasil pengujian yang telah dilakukan pada catbot dengan berbagai parameter uji, sebagai berikut:

Tabel 1. Rekap Hasil Pengujian

Parameter Uji	**Metrik**	**Nilai Rata-rata**
User Acceptance Test (UAT)	Kepuasan Pengguna	88%
Latensi Respons	Kecepatan jawaban dikirim	2.8 Detik
Success Rate	Keberhasilan menarik data tepat	95%
Cost Efficiency	Biaya per 1.000 pesan	< $1.50 (API gpt-4o-mini)

Data rekap pengujian yang telah dilakuakan yang tertera pada Tabel 1, menunjukkan bahwa nilai latensi rata-rata sebesar 2,8 detik tidak hanya dipengaruhi oleh satu komponen, tetapi merupakan hasil akumulasi dari beberapa tahapan proses di dalam sistem. Waktu tersebut mencakup proses pengolahan input, pencarian data (*retrieval*), hingga pemanggilan API OpenAI untuk menghasilkan respons. Di sisi lain, nilai *success rate* sebesar 95% menunjukkan bahwa sistem mampu mengambil dan memanfaatkan data referensi dengan tingkat ketepatan yang tinggi. Hal ini mengindikasikan bahwa mekanisme RAG yang diterapkan pada workflow n8n dapat bekerja dengan cukup konsisten dalam menghubungkan pertanyaan pengguna dengan dokumen yang relevan. Pemisahan sumber data ke dalam beberapa kategori seperti telinga, hidung, dan tenggorokan, turut membantu sistem dalam mempersempit ruang pencarian sehingga konteks yang diperoleh menjadi lebih tepat sasaran.

Penerapan RAG yang diimplementasikan menggunakan platform n8n dengan memanfaatkan mekanisme pemanggilan dokumen secara langsung melalui *tools* tanpa penggunaan *vector store*. Meskipun lebih sederhana, hasil pengujian menunjukkan bahwa sistem tetap mampu mencapai nilai *success rate* sebesar 95%. Hal ini menunjukkan bahwa pemetaan pertanyaan pengguna terhadap sumber data masih dapat dilakukan secara efektif, terutama karena struktur data yang digunakan sudah terorganisir berdasarkan kategori. Jika dibandingkan dengan pendekatan RAG konvensional, perbedaan utama terletak pada kompleksitas sistem. RAG berbasis n8n yang digunakan dalam penelitian ini tidak memerlukan proses embedding maupun penyimpanan vektor, sehingga lebih ringan dari sisi implementasi. Namun demikian, dari sisi hasil, tingkat keberhasilan yang diperoleh tidak menunjukkan penurunan yang signifikan untuk skala data yang digunakan. Dari sisi kinerja, latensi rata-rata sebesar 2,8 detik menunjukkan bahwa sistem masih berada dalam batas respons yang dapat diterima. Nilai ini mencakup seluruh proses, mulai dari penerimaan input, pemanggilan data melalui *tools*, hingga proses generasi oleh model bahasa. Jika dibandingkan dengan sistem RAG berbasis arsitektur kompleks seperti pada penelitian yang telah dilakukan [16][17], perbedaan waktu respons tidak terlalu mencolok, sehingga pendekatan ini dapat dianggap cukup efisien. Dengan demikian, dapat dilihat bahwa implementasi RAG menggunakan n8n memberikan alternatif yang lebih sederhana tanpa menghilangkan fungsi utama dari arsitektur RAG itu sendiri. Pendekatan ini menunjukkan bahwa sistem berbasis RAG tidak selalu harus dibangun dengan infrastruktur yang kompleks, tetapi tetap dapat diimplementasikan secara praktis dengan hasil yang tetap relevan terhadap kebutuhan pengguna.

4. Kesimpulan

Penelitian ini menunjukkan bahwa sistem chatbot berbasis Retrieval-Augmented Generation (RAG) dapat diimplementasikan dengan baik melalui pemanfaatan platform n8n yang terintegrasi dengan WhatsApp sebagai media komunikasi. Selain itu, penggunaan platform *low-code* seperti n8n memberikan keuntungan dalam hal efisiensi pengembangan. Proses perancangan hingga implementasi sistem dapat dilakukan dalam waktu yang relatif singkat tanpa memerlukan pengkodean yang kompleks.

Berdasarkan hasil pengujian, sistem mampu memberikan respons dengan waktu yang masih dapat diterima serta tingkat kesesuaian jawaban yang tinggi terhadap data referensi. Hal ini menunjukkan bahwa integrasi antara mekanisme *retrieval* dan model generatif dapat berjalan secara selaras dalam satu alur kerja yang terstruktur. Di sisi lain, penggunaan WhatsApp sebagai antarmuka juga berkontribusi terhadap kemudahan akses, karena pengguna tidak perlu beradaptasi dengan platform baru. Meskipun demikian, terdapat beberapa hal yang masih dapat dikembangkan lebih lanjut. Pengelolaan data dalam skala yang lebih besar berpotensi membutuhkan pendekatan yang lebih optimal, seperti pemanfaatan *vector database* yang bersifat persisten agar proses pencarian dapat dilakukan dengan lebih efisien. Selain itu, penambahan fitur analisis terhadap isi percakapan, seperti identifikasi tingkat urgensi atau sentimen pengguna, dapat membantu sistem dalam memberikan prioritas terhadap kondisi tertentu.

Secara keseluruhan, hasil penelitian ini memberikan gambaran bahwa pendekatan RAG yang dikombinasikan dengan platform *low-code* dan aplikasi pesan instan dapat menjadi alternatif yang cukup praktis dalam pengembangan sistem chatbot. Pendekatan ini tidak hanya relevan dari sisi teknis, tetapi juga memiliki potensi untuk diadaptasi lebih luas pada berbagai kebutuhan layanan berbasis informasi.

5. Daftar Pustaka

[1] N. Lelyana, “Analisis Dampak Inovasi Teknologi pada Strategi Manajemen Rumah Sakit,” *JISHUM (Jurnal Ilmu Sos. dan Humaniora)*, vol. 2, no. 4, pp. 425–446, 2024, [Online]. Available: https://journal.ikmedia.id/index.php/jishum%0AVol.

[2] I. A. Nurcahyani, “Hubungan Teknologi dan Organisasi dengan Kepuasaan Pengguna dalam Penerapan Sistem Informasi Manajemen Rumah Sakit (SIMRS) di Rumah Sakit Umum Daerah

Ajibarang," *J. Manaj. Inf. Kesehat. …*, vol. 12, no. 1, pp. 90–95, 2024, [Online]. Available: https://jmiki.aptirmik.or.id/jmiki/article/view/653

[3] M. Mustoha *et al.*, "Implementasi Layanan Berbasis Teknologi Informasi Dalam Mewujudkan Pelayanan Unggulan di Rumah Sakit," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, pp. 1915–1921, 2025, doi: 10.31004/riggs.v4i3.2251.

[4] H. S. Yudha Adi Pradana Djatioetomo, Deviana, "Karakteristik Penyakit Pada Poli Tht-Kl Di Rs Mitra Sehat Medika , Pandaan," *Malang Otorhinolaryngoloy Head Neck Surg. J.*, 2022, [Online]. Available: http://moj.ub.ac.id/

[5] N. Rahmayanti, U. Halimatu Sa'diyah, R. Widianto Sudjud, and V. Paramarta, "Penerapan Sistem Informasi Rumah Sakit dalam Meningkatkan Efisiensi Pelayanan di Rumah Sakit," *COMSERVA J. Penelit. dan Pengabdi. Masy.*, vol. 3, no. 08, pp. 3094–3101, 2023, doi: 10.59141/comserva.v3i08.1094.

[6] Y. S. Pongtambing and E. S. Manapa, "Sistem Informasi Kesehatan Dan Telemedicine : Narrative Review informasi kesehatan (SIK) adalah penerimaan pengguna sistem seperti Telemedicine oleh profesional kesehatan dengan menggunakan teknologi informasi dan komunikasi . Pelayanan," vol. 1, no. 4, 2023.

[7] R. Daka and B. P. Jatusari, "Transformasi Digital Kesehatan: Manfaat Kecerdasan Buatan (AI) dalam Pelayanan Kesehatan Bagi Penyandang Disabilitas (Tinjauan Sistematis)," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, pp. 6326–6336, 2025, doi: 10.31004/riggs.v4i3.2926.

[8] D. Abror and Rousyati, "Etika Dan Bias Dalam Llm: Tanggung Jawab Sosial Atas Kecerdasan Buatan Generatif," *J. Unitek*, vol. 18, no. 1, pp. 69–75, 2025, doi: 10.52072/unitek.v18i1.1386.

[9] T. Q. Ramadhani, N. Q. Nada, and N. D. S, "Penerapan Metode Retrieval-Augmented Generation (RAG) Pada Chatbot E-Commerce Berbasis Gemini Ai," *J. Ilm. Ilk. - Ilmu Komput. Inform.*, vol. 8, no. 2, pp. 301–313, 2025, doi: 10.47324/ilkominfo.v8i2.384.

[10] M. A. Basallamah, L. S. Riza, and A. Anisyah, "Pengembangan Chatbot Informasi Kesehatan Ibu dan Anak Jawa Barat Berbasis Hybrid RAG dan TextToSQL," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 3, pp. 1895–1902, 2026, doi: 10.62712/juktisi.v4i3.783.

[11] I. P. P. Tanaya, T. N. Fatyanosa, and H. F. Putra, "Penerapan Synthetic Context Generation Menggunakan Large Language Model pada Sistem Question Answering Berbasis Retrieval-Augmented Generation untuk Domain Kesehatan Gizi," vol. 10, no. 1, pp. 1–10, 2026.

[12] D. Kurniawan and J. Triloka, "Penerapan Teknologi Langchain dan LLM pada Sistem Question Answering Berbasis Chatbot Telegram: Literature Review," *Semin. Nas. Has. Penelit. dan Pengabdi. Masy. 2025*, pp. 95–104, 2025.

[13] M. R. Rachman, M. Rosidin, and W. Y. Sulistyo, "Implementasi Metode Retrieval Augmented Generation Pada Chatbot Untuk Otomatisasi Layanan Pelanggan Kontrakan," *J. Tek. Inform.*, vol. 11, no. 02, p. 229, 2025.

[14] E. S. 1, S. Hidayat2, and Sofwandi Noor, "IMPLEMENTASI N8N UNTUK ANALISIS DATA JARINGAN TELCO BERBASIS OTOMASI WORKFLOW," vol. 13, no. 01, pp. 58–65, 2026.

[15] Aliyah Aliyah, Nahrun Hartono, and Asrul Azhari Muin, "Penggunaan User Acceptance Testing (UAT) Pada Pengujian Sistem Informasi Pengelolaan Keuangan Dan Inventaris Barang," *Switch J. Sains dan Teknol. Inf.*, vol. 3, no. 1, pp. 84–100, 2024, doi: 10.62951/switch.v3i1.330.

[16] S. Aliphadji Talaohu, R. Soekarta, and M. Surahmanto, "Implementasi LLM Pada Chatbot PMB Universitas Muhammadiyah Sorong Menggunakan Metode RAG Berbasis Website," *J. Ilmu Komput. dan Inform.*, vol. 03, no. 02, pp. 1–11, 2025.

[17] A. Farisi, C. Chriestina, D. Dafid, and A. Jansa, "Penerapan Model AI Gemini pada Pengembangan

Chatbot Tanya Fiqih Haji dan Umrah," *INSOLOGI J. Sains dan Teknol.*, vol. 4, no. 6, pp. 1778–1787, 2025, doi: 10.55123/insologi.v4i6.7379.