Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer
Volume 6, Nomor 1, Maret 2026
e-ISSN: 2827-7953; p-ISSN: 2827-8135, Hal. 538-549
DOI: https://doi.org/10.55606/juisik.v6i1.2240
Tersedia: https://journal.sinov.id/index.php/juisik

Implementasi Sistem *Chatbot* Kesehatan Berbasis *Retrieval-Augmented Generation* (RAG) dengan Dataset Medis Bahasa Indonesia

Gusti Ayu Purna Savitri[1*], Adie Wahyudi Oktavia Gama[2]
[1-2] Teknologi Informasi, Universitas Pendidikan Nasional, Indonesia
purnasavitri@gmail.com[1], adiewahyudi@undiknas.ac.id[2]
**Penulis Korespondensi: purnasavitri@gmail.com*

Abstract. *The availability of accurate and easy-to-understand health information in Indonesian remains a significant challenge in the digital era. People tend to rely on unverified sources of information, potentially fueling the spread of health misinformation. This research aims to develop a Retrieval-Augmented Generation (RAG)-based health chatbot capable of providing structured medical responses with traceable references. The system is implemented using the large Qwen 2.5-7B-Instruct language model, the FAISS vector index, and a dataset containing several health questions and answers in Indonesian. The architecture is designed to understand natural language health queries, generate evidence-based responses, and include source links for independent verification. Testing results show that the system successfully answers common health questions by integrating trusted sources, implementing guardrail mechanisms in the form of clinical disclaimers and query filters in external domains, and achieving adequate response times for its initial health information assistant function. This system has been deployed as a web application and has the potential for further development as a component of Indonesia's digital health ecosystem to improve public health literacy and reduce reliance on non-medical information.*

Keywords: *Digital Health; Expert System; Health Chatbot; Indonesian; Retrieval-Augmented Generation.*

Abstrak. Ketersediaan informasi kesehatan yang akurat dan mudah dipahami dalam Bahasa Indonesia masih menjadi tantangan signifikan di era digital. Masyarakat cenderung mengandalkan sumber informasi yang tidak terverifikasi, sehingga berpotensi memicu penyebaran misinformasi kesehatan. Penelitian ini bertujuan mengembangkan chatbot kesehatan berbasis Retrieval-Augmented Generation (RAG) yang mampu memberikan respons medis terstruktur dengan referensi yang dapat dilacak. Sistem diimplementasikan menggunakan model bahasa besar Qwen 2.5-7B-Instruct, indeks vektor FAISS, dan dataset yang berisi pasangan pertanyaan-jawaban kesehatan dalam Bahasa Indonesia. Arsitektur dirancang untuk memahami kueri kesehatan dalam bahasa alami, menghasilkan respons berbasis bukti, serta menyertakan tautan sumber untuk verifikasi independen. Hasil pengujian menunjukkan bahwa sistem berhasil menjawab pertanyaan kesehatan umum dengan mengintegrasikan sumber terpercaya, menerapkan mekanisme guardrail berupa disclaimer klinis dan penyaring kueri di luar domain, serta mencapai waktu respons yang memadai untuk fungsi asisten informasi kesehatan awal. Sistem ini telah di-deploy sebagai aplikasi web dan berpotensi dikembangkan lebih lanjut sebagai komponen ekosistem kesehatan digital Indonesia guna meningkatkan literasi kesehatan masyarakat dan mengurangi ketergantungan pada informasi non-medis.

Kata Kunci: Bahasa Indonesia; *Chatbot* Kesehatan; Kesehatan Digital; *Retrieval-Augmented Generation*; Sistem Pakar.

1. PENDAHULUAN

Di era digital, masyarakat Indonesia semakin bergantung pada internet sebagai sumber utama informasi kesehatan. Tingginya penetrasi internet dan penggunaan perangkat seluler telah mengubah cara masyarakat mencari informasi terkait gejala penyakit, pilihan pengobatan, hingga langkah pencegahan (Swacha & Gracel, 2025). Fenomena ini mencerminkan meningkatnya kesadaran masyarakat akan pentingnya kesehatan, namun sekaligus menimbulkan kekhawatiran serius terkait kualitas, validitas, dan keamanan informasi yang dikonsumsi.

Naskah Masuk: 31 Januari 2026; Revisi: 24 Februari 2026; Diterima: 28 Maret 2026; Terbit: 31 Maret 2026

Permasalahan utama yang dihadapi adalah keterbatasan ketersediaan literatur medis yang akurat, komprehensif, dan mudah dipahami dalam Bahasa Indonesia. Sebagian besar panduan klinis dan jurnal bereputasi masih tersedia dalam Bahasa Inggris, sehingga menciptakan hambatan bahasa bagi masyarakat umum (Firdaus et al., 2024). Akibatnya, banyak individu beralih ke sumber informasi tidak terverifikasi seperti media sosial, blog pribadi, atau forum diskusi tanpa pengawasan tenaga medis. Kondisi ini berpotensi memicu misinformasi yang dapat berakibat fatal, mulai dari diagnosis mandiri yang keliru, penggunaan obat tanpa indikasi, hingga penundaan pencarian pertolongan medis profesional (Zhang et al., 2026; Coric et al., 2026).

Teknologi chatbot berbasis kecerdasan buatan menawarkan solusi potensial untuk menjembatani kesenjangan tersebut. Chatbot kesehatan dapat memberikan respons cepat terhadap pertanyaan kesehatan dasar, tersedia 24/7, dan menjangkau pengguna secara masif (Valan & Venugopal, 2025; Nayinzira & Adda, 2024). Beberapa penelitian terdahulu telah mengembangkan chatbot kesehatan dengan pendekatan berbasis aturan, machine learning, atau deep learning. Namun, model konvensional sering menghasilkan jawaban yang terlalu umum, rentan terhadap halusinasi akibat keterbatasan pengetahuan parametrik, serta tidak menyertakan referensi yang dapat diverifikasi (Muhetaer et al., 2025; Patil et al., 2025). Selain itu, sistem tersebut umumnya kesulitan menangani kueri yang berada di luar distribusi data pelatihan (Xu et al., 2024).

Pendekatan *Retrieval-Augmented Generation* (RAG) muncul sebagai solusi inovatif untuk meningkatkan akurasi dan transparansi chatbot medis (Sohn, 2024; Bora & Cuayáhuitl, 2024). RAG mengintegrasikan mekanisme pengambilan informasi dari basis pengetahuan eksternal dengan kemampuan generatif model bahasa besar. Dengan arsitektur ini, sistem tidak hanya mengandalkan pengetahuan yang tertanam selama pelatihan, tetapi secara aktif mengambil dan merujuk pada dokumen sumber yang relevan sebelum menyusun respons. Keunggulan utama RAG meliputi kemampuan menyediakan jawaban berbasis bukti dengan sitasi eksplisit, mengurangi risiko halusinasi, serta memungkinkan pembaruan pengetahuan secara dinamis tanpa memerlukan pelatihan ulang model (Haider, 2025; Miao et al., 2024). Pendekatan ini juga meningkatkan akuntabilitas sistem, yang sangat krusial dalam konteks pengambilan keputusan kesehatan (Baur et al., 2025).

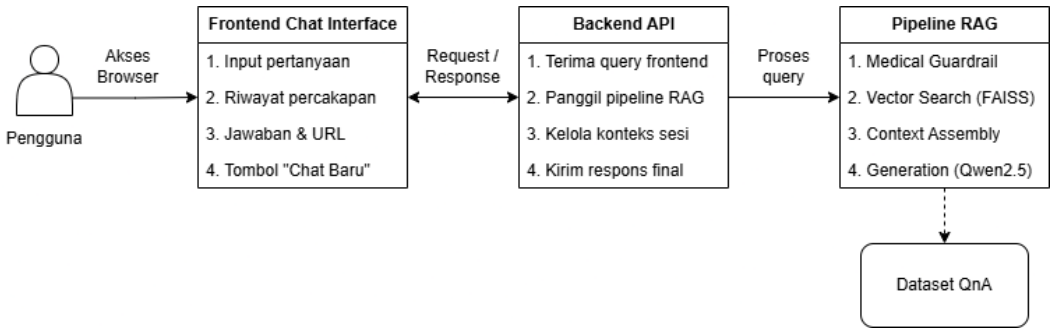
Dalam konteks Indonesia, pengembangan *chatbot* kesehatan berbasis RAG menghadapi tantangan khusus, antara lain ketersediaan dataset medis berkualitas dalam Bahasa Indonesia yang masih terbatas, kebutuhan model untuk memahami nuansa bahasa informal, serta kepatuhan terhadap standar etika dan privasi data kesehatan (Meng et al., 2025; Benfenati et

al., 2024). Penelitian ini bertujuan mengembangkan chatbot kesehatan berbasis RAG yang memanfaatkan dataset Q&A kesehatan sebagai sumber pengetahuan terverifikasi. Fokus penelitian mencakup implementasi pipeline pemrosesan yang mencakup penyaringan domain, pencarian vektor semantik, dan generasi respons terstruktur, serta evaluasi komprehensif terhadap aspek keandalan, relevansi, dan keamanan sistem (Long, 2024; Shin et al., 2025). Diharapkan sistem ini dapat menjadi pondasi awal dalam penyediaan akses informasi kesehatan yang terpercaya, mengurangi misinformasi, dan mendukung peran tenaga medis dalam edukasi masyarakat (Ziletti & DAmbrosi, 2024).

2. METODE

Arsitektur Sistem

Sistem chatbot kesehatan dikembangkan dengan mengadopsi arsitektur tiga lapis yang terdiri dari antarmuka pengguna (*frontend*), lapisan logika aplikasi (*backend*), dan alur pemrosesan *Retrieval-Augmented Generation* (RAG). Arsitektur sistem secara lengkap ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Sistem Chatbot Kesehatan Berbasis RAG.

Pada lapisan *frontend*, pengguna berinteraksi melalui antarmuka berbasis web yang memungkinkan input pertanyaan kesehatan, menampilkan riwayat percakapan, serta menyajikan jawaban beserta tautan referensi sumber. Fitur utama antarmuka meliputi kolom input pertanyaan, tampilan riwayat percakapan, jawaban dengan sumber referensi, dan tombol "Percakapan Baru" untuk memulai sesi baru. Lapisan backend berfungsi sebagai orkestrator yang menerima permintaan pengguna, mengelola konteks sesi, memanggil modul RAG, serta mengembalikan respons terformat (Son et al., 2025). *Pipeline* RAG memproses *query* melalui empat tahap utama: (1) *Medical Guardrail* untuk menyaring pertanyaan non-medis, (2) *Vector Search* menggunakan FAISS untuk mencari dokumen relevan, (3) *Context Assembly* untuk mengumpulkan konteks jawaban, dan (4) *Generation* menggunakan Qwen 2.5 untuk menghasilkan respons akhir.

Pengumpulan dan Preprocessing Data

Dataset utama yang digunakan dalam penelitian ini berasal dari kumpulan tanya jawab kesehatan Alodokter yang tersedia secara publik melalui platform *Hugging Face*. Dataset ini mengandung pasangan pertanyaan-jawaban yang telah ditinjau oleh tenaga medis, sehingga menjamin validitas klinis dan relevansi kontekstual. Struktur dataset ditunjukkan pada Tabel1.

Tabel 1. Struktur Dataset Alodokter Q&A.

Kolom	Deskripsi
title	Judul atau *headline* diskusi kesehatan
question	Pertanyaan detail dari pengguna/pasien dalam bahasa Indonesia alami
answer	Jawaban lengkap dari dokter
doctor_name	Nama dokter yang memberikan jawaban
tag	Tag atau kategori topik kesehatan terkait
url	Tautan sumber asli untuk transparansi dan verifikasi

Sebelum diintegrasikan ke dalam pipeline, data melalui tahap pembersihan untuk menghapus konten redundan, salam generik, dan informasi sensitif. Selanjutnya, teks dipotong (*chunking*) pada level kalimat dengan panjang minimum tertentu guna mempertahankan koherensi semantik selama proses pengambilan informasi (Benfenati et al., 2024).

Implementasi *Retrieval-Augmented Generation*

Komponen awal implementasi RAG adalah representasi vektor dan mekanisme pencarian semantik. Model embedding multibahasa dipilih untuk mengonversi kueri pengguna dan dokumen referensi menjadi representasi vektor padat, dengan pertimbangan utama pada dukungan optimal terhadap struktur linguistik Bahasa Indonesia. Vektor yang dihasilkan diindeks menggunakan FAISS, memungkinkan pencarian dokumen relevan secara efisien berdasarkan metrik cosine similarity. Pendekatan ini memastikan bahwa dokumen yang diambil memiliki relevansi semantik tinggi terhadap intent pengguna sebelum diproses lebih lanjut (Bora & Cuayáhuitl, 2024).

Pada tahap generasi, sistem memanfaatkan model instruktif yang telah dioptimalkan untuk tugas tanya jawab. Qwen 2.5-7B-Instruct dipilih karena kemampuannya mengikuti instruksi kompleks dan menghasilkan teks koheren secara multibahasa. Untuk menyeimbangkan kualitas respons dengan efisiensi komputasi, teknik kuantisasi diterapkan agar model dapat *di-deploy* pada lingkungan dengan sumber daya terbatas tanpa penurunan performa signifikan (Haider, 2025).

Integrasi antara modul retrieval dan generation diatur melalui rekayasa prompt yang terstruktur. Instruksi dalam prompt mengarahkan model untuk menyusun respons yang sepenuhnya berlandaskan pada konteks dokumen yang diambil, sehingga meminimalkan risiko halusinasi. Selain itu, prompt juga menetapkan standar bahasa Indonesia yang profesional,

mudah dipahami, serta mewajibkan penyertaan referensi sumber pada setiap respons guna meningkatkan transparansi dan akuntabilitas sistem (Miao et al., 2024).

Evaluasi

Evaluasi sistem menggunakan pendekatan LLM-as-a-Judge, yang memanfaatkan model bahasa besar independen sebagai penilai otomatis untuk mengukur kualitas respons yang dihasilkan. Pendekatan ini dipilih karena kemampuannya menilai aspek kompleks seperti konsistensi semantik, kesesuaian konteks, dan keamanan respons secara objektif dan terukur dibandingkan evaluasi manual konvensional. Model penilai yang digunakan adalah Qwen 2.5-14B-Instruct, dipilih berbeda dari model utama sistem untuk menghindari bias evaluasi dan menjamin independensi hasil.

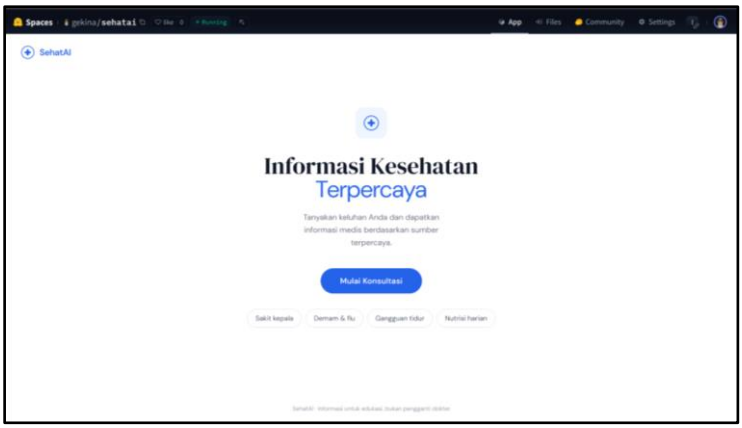
Penilaian dilakukan terhadap lima dimensi utama yang merefleksikan kualitas dan kinerja sistem secara holistik. Aspek *Faithfulness* mengukur sejauh mana respons hanya bersumber dari konteks dokumen yang diambil, sehingga meminimalkan halusinasi. Aspek *Context Precision* menilai keterkaitan respons dengan maksud kueri pengguna. Aspek *Correctness* mengevaluasi keakuratan informasi medis dibandingkan sumber referensi terpercaya. Aspek *Refusal Rate* mengukur kemampuan sistem mengenali dan menolak kueri di luar domain kesehatan atau yang berpotensi membahayakan. Terakhir, *aspek Latency* diukur sebagai indikator kinerja teknis sistem dalam menyediakan respons.

Proses evaluasi dilaksanakan dengan merancang prompt penilaian terstruktur yang mengarahkan model penilai untuk memberikan skor berdasarkan kriteria baku. Untuk *Faithfulness*, *Relevance*, *Correctness*, dan *Refusal Rate*, penilaian menggunakan skala biner (0/1) yang diagregasikan menjadi persentase keberhasilan, sedangkan *Correctness* juga dapat dinilai menggunakan skala Likert 1–5. Prompt dirancang dalam Bahasa Indonesia untuk memastikan pemahaman konteks optimal dan telah diuji coba secara manual sebelum diterapkan pada seluruh dataset uji.

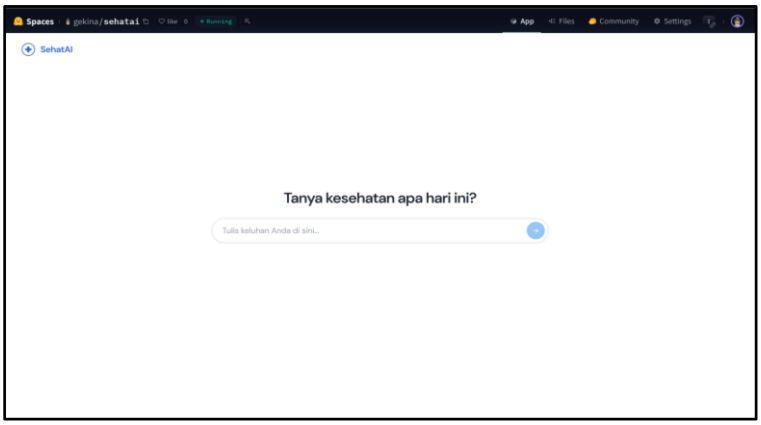
3. HASIL DAN PEMBAHASAN

Tampilan Antarmuka Sistem

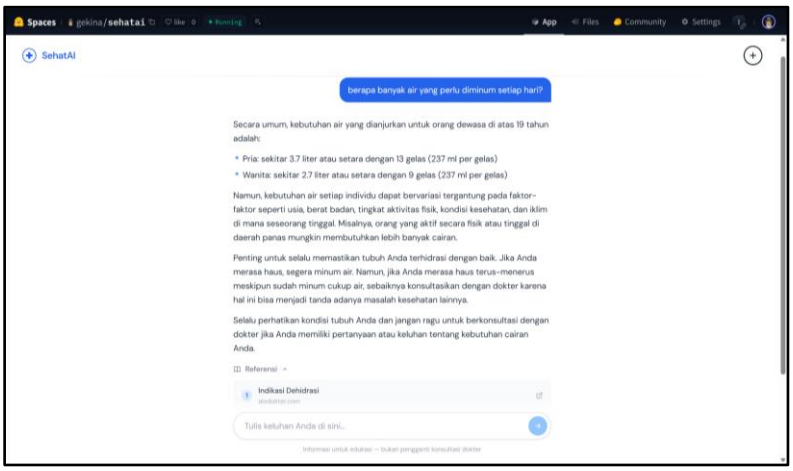
Sistem chatbot kesehatan berhasil diimplementasikan dengan antarmuka web yang dirancang intuitif untuk memudahkan pengguna dari berbagai latar belakang. Tampilan utama sistem disajikan pada Gambar 2, Gambar 3, dan Gambar 4.



Gambar 2. Tampilan Home.



Gambar 3. Tampilan Input Awal.

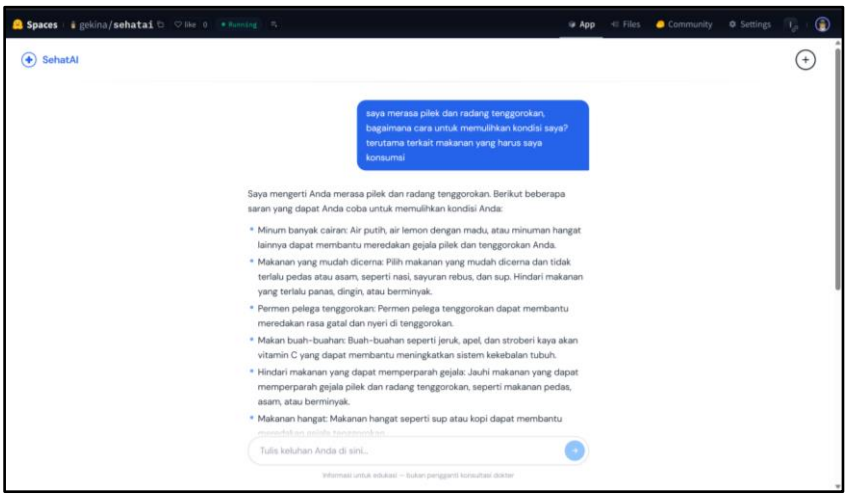


Gambar 4. Tampilan Chat.

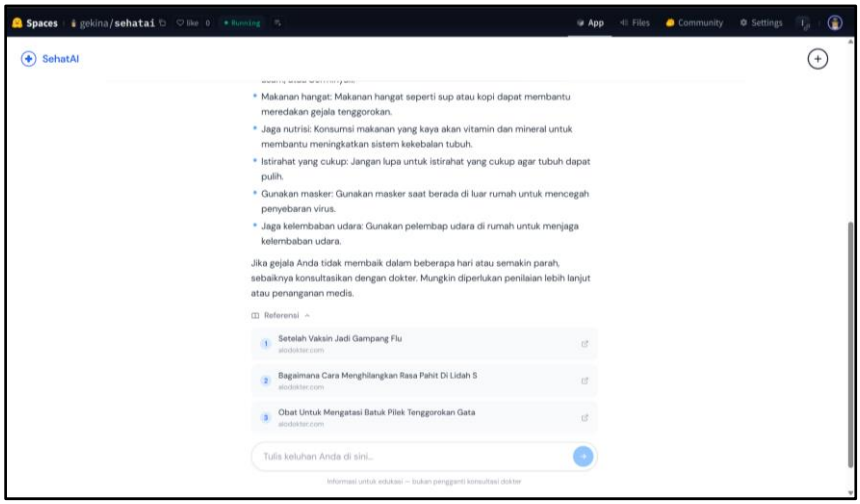
Antarmuka terdiri dari halaman landing yang menekankan kredibilitas ("Informasi Kesehatan Terpercaya") serta menyediakan akses cepat melalui empat tombol kategori keluhan umum. Setelah pengguna menekan "Mulai Konsultasi", pengguna akan diarahkan ke tampilan awal dengan input box yang ada di bagian tengah. Setelah pengguna memberikan pertanyaan, selanjutnya sistem beralih ke antarmuka percakapan yang menampilkan riwayat interaksi dalam format *bubble chat*, kolom input responsif, serta tombol + untuk memulai sesi baru.

Setiap jawaban AI secara otomatis menyertakan panel "Referensi" yang dapat dibuka-tutup, menampilkan daftar tautan ke artikel Alodokter yang menjadi dasar penyusunan respons. Desain antarmuka mengutamakan keterbacaan, aksesibilitas, dan transparansi sumber, serta dilengkapi disclaimer klinis di bagian bawah layar sebagai mekanisme guardrail pasif yang mengingatkan pengguna bahwa sistem bersifat edukatif dan tidak menggantikan konsultasi profesional.

Contoh Hasil Jawaban Chatbot



Gambar 5. Contoh Hasil Jawaban Chatbot.



Gambar 6. Contoh Hasil Jawaban Chatbot + Referensi.

Hasil menunjukkan bahwa sistem mampu memahami pertanyaan dalam Bahasa Indonesia alami yang mengandung jawaban yang disusun berdasarkan referensi dan menghasilkan respons yang relevan secara klinis. Setiap jawaban secara konsisten menyertakan tautan ke sumber asli dari dataset Alodokter melalui panel referensi, sehingga memungkinkan pengguna menelusuri informasi lebih lanjut dan memverifikasi keakuratan respons yang diberikan (Nayinzira & Adda, 2024).

Kinerja Sistem

Penilaian dilaksanakan menggunakan pendekatan LLM-as-a-Judge dengan lima metrik utama: *Faithfulness, Relevance, Correctness, Refusal Rate*, dan *Latency*. Hasil evaluasi dirangkum pada Tabel 4.

Tabel 4. Metrik Kinerja Sistem.

Metrik	Nilai
Faithfulness	72%
Context Precision	74%
Correctness	3.58
Refusal Rate	2%
Latency	45.58 s

Sistem menunjukkan kinerja yang memuaskan pada aspek Faithfulness dengan nilai 72%, yang mengindikasikan bahwa sebagian besar respons hanya bersumber dari konteks dokumen yang diambil dan meminimalkan risiko halusinasi. Aspek Context Precision mencatat nilai 74%, menunjukkan bahwa respons umumnya selaras dengan maksud dan konteks pertanyaan pengguna. Pada aspek Correctness, sistem memperoleh skor rata-rata 3.58 pada skala 1–5, yang merefleksikan tingkat keakuratan informasi medis berdasarkan sumber terpercaya (Zhang et al., 2026).

Mekanisme guardrail menunjukkan efektivitas tinggi dengan Refusal Rate sebesar 2%, artinya sistem mampu mengenali dan menolak pertanyaan di luar domain kesehatan atau yang berpotensi membahayakan, sambil tetap memberikan panduan yang sopan. Sementara itu, waktu respons (Latency) tercatat rata-rata 45.58 detik, yang dipengaruhi oleh proses embedding, pencarian vektor, dan generasi respons secara *real-time*. Meskipun nilai ini masih dapat dioptimasi, waktu respons dinilai memadai untuk penggunaan sebagai asisten informasi kesehatan awal yang tidak memerlukan interaksi seketika.

Pembahasan

Secara keseluruhan, sistem chatbot kesehatan berbasis RAG berhasil memenuhi tujuan penelitian, yaitu menyediakan respons medis dengan referensi yang jelas dan dapat diverifikasi. Keunggulan utama sistem terletak pada transparansi sumber informasi, yang merupakan aspek krusial dalam konteks kesehatan untuk membangun kepercayaan pengguna dan mengurangi risiko misinformasi. Integrasi mekanisme guardrail (baik melalui disclaimer UI maupun penyaring kueri) juga terbukti efektif dalam menjaga fokus sistem pada domain kesehatan dan mencegah pemberian saran yang berpotensi berbahaya (Baur et al., 2025).

Namun, beberapa keterbatasan teridentifikasi selama pengujian. Pertama, latensi sistem masih relatif tinggi untuk skenario yang memerlukan respons instan. Kedua, kualitas respons sangat bergantung pada kelengkapan dan cakupan dataset sumber; pertanyaan yang sangat spesifik atau jarang muncul dalam dataset mungkin tidak terjawab secara optimal. Ketiga, sistem saat ini hanya mendukung Bahasa Indonesia sehingga belum dapat menjangkau pengguna yang menggunakan bahasa daerah atau Bahasa Inggris.

Dibandingkan dengan chatbot kesehatan konvensional, sistem berbasis RAG menawarkan keunggulan signifikan dalam hal akurasi dan akuntabilitas informasi. Kemampuan merujuk pada sumber eksternal memungkinkan sistem untuk tetap mutakhir tanpa memerlukan pelatihan ulang model, serta memberikan jejak verifikasi yang jelas bagi pengguna. Untuk pengembangan selanjutnya, optimasi dapat dilakukan melalui teknik reranking dokumen, fine-tuning model embedding untuk domain kesehatan Indonesia, penambahan sumber data dari institusi kesehatan terpercaya, serta integrasi dengan layanan telemedicine untuk rujukan konsultasi profesional.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan chatbot kesehatan berbasis Retrieval-Augmented Generation (RAG) yang memanfaatkan sumber informasi medis Bahasa Indonesia dari dataset Alodokter Q&A. Sistem mengintegrasikan model embedding multibahasa, indeks vektor FAISS, dan Qwen 2.5-7B-Instruct untuk menghasilkan respons kesehatan yang relevan dan berbasis bukti. Evaluasi menggunakan pendekatan LLM-as-a-Judge menunjukkan bahwa chatbot mampu memberikan jawaban dengan tingkat kesetiaan terhadap konteks sumber sebesar 72%, relevansi 74%, dan keakuratan informasi medis yang memadai. Sistem juga mencatat tingkat penolakan yang tinggi terhadap pertanyaan di luar domain medis, dengan waktu respons rata-rata 45.58 detik yang dinilai memadai untuk fungsi asisten informasi kesehatan awal.

Keunggulan utama sistem terletak pada transparansi informasi, di mana setiap respons dilengkapi dengan panel referensi yang mengarah ke sumber asli terverifikasi. Pendekatan RAG memungkinkan sistem mengurangi risiko halusinasi dan memudahkan pembaruan pengetahuan tanpa pelatihan ulang yang kompleks. Dengan antarmuka web yang sederhana dan fitur akses cepat untuk keluhan umum, sistem ini berpotensi dikembangkan sebagai asisten informasi kesehatan awal yang menjangkau masyarakat Indonesia, khususnya dalam upaya mengurangi penyebaran misinformasi kesehatan.

Meskipun demikian, sistem masih memiliki keterbatasan yang perlu diatasi, antara lain latensi komputasi real-time, ketergantungan pada kelengkapan dataset, serta dukungan bahasa yang terbatas. Penelitian mendatang dapat difokuskan pada implementasi teknik caching atau pemrosesan asinkron untuk meningkatkan kecepatan respons, penambahan sumber data dari institusi kesehatan terverifikasi, serta integrasi dengan platform telemedicine untuk memfasilitasi rujukan profesional. Dengan penyempurnaan tersebut, sistem ini diharapkan dapat menjadi komponen integral dalam ekosistem kesehatan digital Indonesia yang terpercaya dan berorientasi pada kebutuhan masyarakat.

DAFTAR PUSTAKA

Baur, D., Ansorg, J., Heyde, C.-E., & Voelker, A. (2025). Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. *JMIR AI*, *4*, e75262. https://doi.org/10.2196/75262

Benfenati, D., De Filippis, G. M., Rinaldi, A. M., Russo, C., & Tommasino, C. (2024). A Retrieval-augmented Generation application for Question-Answering in Nutrigenetics Domain. *Procedia Computer Science*, *246*, 586–595. https://doi.org/10.1016/j.procs.2024.09.467

Bora, A., & Cuayáhuitl, H. (2024). Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Machine Learning and Knowledge Extraction (MAKE)*, *6*(4), 2355–2374. https://doi.org/10.3390/make6040116

Coric, R., Oloyede, E. F., & Cuayáhuitl, H. (2026). Helpful or Harmful? Re-Evaluating Frugality in Retrieval-Augmented Generation for Medical Question Answering. *Machine Learning and Knowledge Extraction (MAKE)*, *8*(3), 64. https://doi.org/10.3390/make8030064

Firdaus, D., Sumardi, I., & Kulsum, Y. (2024). Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb. *JISKa*, *9*(3), 230–243. https://doi.org/10.14421/jiska.2024.9.3.230-243

Haider, S. A. et al. (2025). The Development and Evaluation of a Retrieval-Augmented Generation Large Language Model Virtual Assistant for Postoperative Instructions. *Bioengineering*, *12*(11), 1219. https://doi.org/10.3390/bioengineering12111219

Long, C. et al. (2024). ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology–Head and Neck Surgery. *Otolaryngology–Head and Neck Surgery*, *171*(4), 1042–1051. https://doi.org/10.1002/ohn.864

Meng, W., Li, Y., Chen, L., & Dong, Z. (2025). Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application. *Electronics*, *14*(2), 386. https://doi.org/10.3390/electronics14020386

Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., & Cheungpasitporn, W. (2024). Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. *Medicina*, *60*(3), 445. https://doi.org/10.3390/medicina60030445

Muhetaer, M., Yusupu, A., Yifan, W., Mutalipu, M., & Hao, F. (2025). Medical QA dialogue datasets in RAG systems performance evaluation and ChatGPT optimization. *Scientific Reports*, *15*(1), 44467. https://doi.org/10.1038/s41598-025-28015-4

Nayinzira, J. P., & Adda, M. (2024). SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis. *Procedia Computer Science*, *251*, 334–341. https://doi.org/10.1016/j.procs.2024.11.118

Patil, R., Abbidi, M., & Fannon, S. (2025). RAGMed: A RAG-Based Medical AI Assistant for Improving Healthcare Delivery. *AI*, *6*(10), 240. https://doi.org/10.3390/ai6100240

Shin, M., Song, J., Kim, M.-G., Yu, H. W., Choe, E. K., & Chai, Y. J. (2025). Thyro-GenAI: A Chatbot Using Retrieval-Augmented Generative Models for Personalized Thyroid Disease Management. *Journal of Clinical Medicine*, *14*(7), 2450. https://doi.org/10.3390/jcm14072450

Sohn, J. et al. (2024). *Rationale-Guided Retrieval Augmented Generation for Medical Question Answering*.

Son, N., Kang, I., Kim, I., Lee, K., Nam, S., & Lee, D. (2025). Development and Evaluation of a Retrieval-Augmented Generation-Based Electronic Medical Record Chatbot System. *Healthcare Informatics Research*, *31*(3), 218–225. https://doi.org/10.4258/hir.2025.31.3.218

Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences*, *15*(8), 4234. https://doi.org/10.3390/app15084234

Valan, P., & Venugopal, P. (2025). Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility–accuracy-readability study of the DIAN AI assistant. *Frontiers in Artificial Intelligence*, *8*, 1640994. https://doi.org/10.3389/frai.2025.1640994

Xu, R., Hong, Y., Zhang, F., & Xu, H. (2024). Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. *Scientific Reports*, *14*(1), 30794. https://doi.org/10.1038/s41598-024-81052-3

Zhang, S., Phan, E., Velmovitsky, P., Pham, Q., & Sanner, S. (2026). Retrieval-Augmented Generation for Medical Question Answering on a Heart Failure Dataset: Performance Analysis. *JMIR Formative Research*, *10*, e84932. https://doi.org/10.2196/84932

Ziletti, A., & DAmbrosi, L. (2024). Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 47–53. https://doi.org/10.18653/v1/2024.clinicalnlp-1.4