JSAI : Journal Scientific and Applied Informatics
Vol. 09, No. 1, *January* 2026, *pages*. 105~111
E-ISSN: 2614-3054; P-ISSN: 2614-3062, *accredited by Kemenristekdikti*, Sinta 4
DOI: 10.36085

Evaluasi *Metode Retrieval* pada *Chatbot Domain* Khusus Berbasis *Retrieval-Augmented Generation*

[1]Asmaidin, [2]Cahyono Budy Santoso
[1,2]Universitas Pembangunan Jaya, Indonesia
[1]asmaidin.2021@student.upj.ac.id; [2]cahyono.budy@upj.ac.id;

Article Info

Article history:

Received, 2026-01-13
Revised, 2026-01-23
Accepted, 2026-01-23

Kata Kunci:

Chatbot
Retrieval-Augmented Generation
evaluasi retrieval
sistem tanya jawab
information retrieval

ABSTRAK

Penelitian ini mengevaluasi metode *retrieval* dalam implementasi *chatbot* domain khusus berbasis *Retrieval-Augmented Generation* untuk meningkatkan akurasi dan relevansi informasi yang disampaikan kepada pengguna serta mengurangi risiko *hallucination*. Permasalahan utama yang diidentifikasi adalah ketidaktepatan pemilihan konteks dan kesalahan prioritas peringkat dokumen pada sistem *chatbot* berbasis model bahasa besar. Penelitian ini menggunakan pendekatan eksperimental dengan membandingkan tiga *metode retrieval*, yaitu pendekatan leksikal berbasis term *frequency–inverse document frequency*, pendekatan semantik berbasis representasi vektor, dan pendekatan *hybrid* yang mengombinasikan keduanya. Evaluasi dilakukan menggunakan metrik *Recall@k* dan *Mean Reciprocal Rank* untuk mengukur kemampuan sistem dalam menemukan dan memprioritaskan dokumen yang relevan. Hasil penelitian menunjukkan bahwa pendekatan leksikal unggul dalam presisi peringkat teratas, sedangkan pendekatan semantik mengalami penurunan kinerja pada domain dokumen teknis. Pendekatan *hybrid* berhasil meningkatkan cakupan dokumen relevan pada peringkat menengah, namun masih menghadapi masalah ambiguïtas peringkat. Temuan ini menegaskan bahwa kualitas sistem *Retrieval-Augmented Generation* sangat bergantung pada mekanisme penentuan prioritas konteks, bukan sekadar kemampuan menemukan informasi.

Keywords:

Chatbot
Retrieval-Augmented Generation
retrieval evaluation
information retrieval question answering systems

ABSTRACT

This study evaluated retrieval methods in the implementation of a domain-specific chatbot based on Retrieval-Augmented Generation to improve information accuracy and relevance while reducing hallucination risks. The primary problem addressed was the incorrect selection and prioritization of contextual documents in chatbot systems built on large language models, particularly in technical domains. An experimental approach was applied by comparing three retrieval strategies: lexical retrieval based on term frequency–inverse document frequency, semantic retrieval using vector representations, and a hybrid retrieval method combining lexical and semantic signals. System performance was measured using Recall at different ranking thresholds and Mean Reciprocal Rank to assess both document discovery and ranking quality. The results demonstrated that lexical retrieval achieved the highest precision at the top-ranked position, while semantic retrieval showed reduced effectiveness due to semantic drift in technical documents. The hybrid approach improved mid-range recall performance but still exhibited ranking ambiguity for top-ranked results. These findings indicated that retrieval quality in Retrieval-Augmented Generation systems depended more on effective ranking and context prioritization than on document availability alone. The study concluded that systematic evaluation of retrieval methods was essential for developing reliable domain-specific chatbots.

Penulis Korespondensi:

Cahyono Budy Santoso,
Universitas Pembangunan Jaya,
Email: cahyono.budy@upj.ac.id

1. PENDAHULUAN

Perkembangan kecerdasan buatan (*Artificial Intelligence*/AI) dalam beberapa tahun terakhir telah mendorong pemanfaatan *chatbot* sebagai antarmuka interaktif untuk sistem informasi di berbagai sektor, khususnya pada domain-domain khusus yang membutuhkan ketepatan dan kedalaman informasi. *Chatbot* tidak lagi berfungsi sekadar sebagai agen percakapan sederhana, tetapi telah berevolusi menjadi sistem cerdas yang mampu mendukung pencarian informasi, pengambilan keputusan, serta pelayanan pengguna secara otomatis dan efisien. Salah satu pendekatan yang banyak digunakan dalam pengembangan *chatbot* modern adalah *Retrieval-Augmented Generation* (RAG). Metode ini mengombinasikan kemampuan *Large Language Models* (LLM) dalam menghasilkan teks dengan mekanisme *retrieval* untuk mengambil informasi relevan dari basis pengetahuan eksternal sebelum proses generasi respons dilakukan. Dengan pendekatan ini, *chatbot* diharapkan mampu memberikan jawaban yang lebih akurat, kontekstual, dan berbasis sumber data yang valid [1][2].

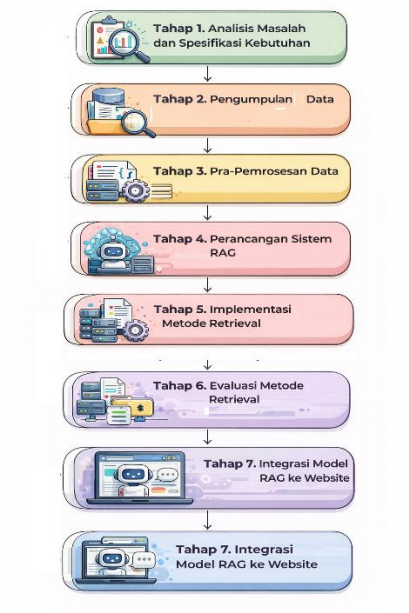
Meskipun demikian, penerapan *chatbot* berbasis *large language models* masih menghadapi tantangan signifikan, terutama fenomena *hallucination*, yaitu kondisi ketika model menghasilkan informasi yang terdengar meyakinkan tetapi tidak sesuai dengan fakta atau sumber yang tersedia [3][4]. Tantangan ini menjadi semakin krusial pada domain khusus seperti kesehatan, pendidikan, dan layanan profesional, di mana kesalahan informasi dapat berdampak serius terhadap pengambilan keputusan dan kepercayaan pengguna. Oleh karena itu, diperlukan evaluasi sistematis terhadap metode *Retrieval-Augmented Generation* (RAG) untuk memastikan bahwa *chatbot* yang dikembangkan mampu menyajikan informasi yang akurat, relevan, dan dapat dipercaya. Kebutuhan akan *chatbot* yang andal juga semakin meningkat pada perusahaan yang bergerak di bidang perdagangan piranti lunak, lisensi aplikasi, dan solusi digital di sektor teknologi informasi, telekomunikasi, serta industri digital, yang meliputi aktivitas pengembangan dan perdagangan perangkat lunak, pemrograman komputer berbasis kecerdasan artifisial, pengembangan aplikasi berbasis internet, penyelenggaraan layanan telekomunikasi, serta instalasi dan reparasi peralatan komunikasi. Dalam konteks tersebut, *chatbot* berperan sebagai media layanan informasi, dukungan teknis, dan komunikasi dengan pengguna, sehingga kualitas retrieval dan ketepatan konteks menjadi faktor kunci dalam menjaga keandalan informasi yang disampaikan [3][4].

Sejumlah penelitian sebelumnya telah mengeksplorasi penggunaan *chatbot* sebagai sistem informasi otomatis berbasis teknik *retrieval*. Rukhiran dan Netinant (2022) menunjukkan bahwa *chatbot* dapat meningkatkan efisiensi penyampaian informasi dengan memanfaatkan mekanisme pengambilan data terstruktur [5]. Penelitian lain mengembangkan *GastroBot*, sebuah *chatbot* medis berbasis RAG yang dirancang khusus untuk menangani informasi penyakit saluran pencernaan, dan menunjukkan bahwa integrasi pengetahuan domain secara eksplisit mampu meningkatkan akurasi respons *chatbot* [1]. Dalam konteks kesehatan, studi oleh [6] memperlihatkan efektivitas *chatbot* berbasis *retrieval* dalam mengelola dan menyederhanakan alur kerja administratif dengan memanfaatkan data dari catatan medis elektronik. Hasil penelitian tersebut menegaskan bahwa pendekatan *retrieval-based* sangat berperan dalam meningkatkan keandalan informasi dan mengurangi beban kerja manusia. Secara keseluruhan, literatur menunjukkan bahwa pendekatan RAG memiliki potensi besar dalam mendukung *chatbot domain* khusus, namun masih diperlukan evaluasi mendalam terkait efektivitasnya dalam mengatasi keterbatasan LLM, khususnya fenomena halusinasi.

Meskipun berbagai penelitian telah menunjukkan keberhasilan chatbot berbasis RAG dalam berbagai domain, masih terdapat celah penelitian yang signifikan terkait evaluasi sistematis *metode retrieval* dan dampaknya terhadap akurasi serta reliabilitas *respons chatbot*. Sebagian besar penelitian berfokus pada implementasi atau studi kasus tertentu, namun belum banyak yang secara eksplisit mengevaluasi bagaimana mekanisme *retrieval* dalam RAG berkontribusi dalam mengurangi fenomena halusinasi dan meningkatkan relevansi informasi [3][7]. Selain itu, evaluasi dari sudut pandang pengguna terhadap efektivitas *chatbot* berbasis RAG masih relatif terbatas. Pemahaman mengenai persepsi pengguna menjadi penting untuk menilai sejauh mana *chatbot* benar-benar mampu memenuhi kebutuhan informasi secara tepat waktu dan akurat. Oleh karena itu, penelitian ini dilakukan untuk mengisi kesenjangan tersebut dengan mengevaluasi metode *retrieval* dan implementasi *chatbot* berbasis RAG secara komprehensif pada konteks domain khusus.

Berdasarkan latar belakang permasalahan dan tinjauan literatur yang telah diuraikan, penelitian ini bertujuan untuk mengevaluasi implementasi dan efektivitas metode *Retrieval-Augmented Generation* (RAG) dalam pengembangan *chatbot* pada *domain* khusus. Penelitian ini secara khusus diarahkan untuk menganalisis peran mekanisme *retrieval* dalam meningkatkan akurasi, relevansi, serta keandalan informasi yang dihasilkan oleh *chatbot* berbasis model bahasa besar. Meskipun pendekatan *Retrieval-Augmented Generation* telah banyak diimplementasikan pada *chatbot* domain khusus, sebagian besar penelitian berfokus pada aspek implementasi sistem, sementara evaluasi sistematis terhadap kualitas *retrieval* dan mekanisme peringkat konteks masih terbatas. Penelitian ini mengisi celah tersebut dengan melakukan evaluasi kuantitatif dan analisis kesalahan terhadap beberapa metode *retrieval* pada sistem RAG.

2. METODE PENELITIAN



Gambar 1 Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang dirancang secara sistematis agar dapat menghasilkan sistem *chatbot* berbasis *Retrieval-Augmented Generation* yang sesuai dengan kebutuhan. Untuk alur penelitian sebagaimana pada gambar 1.

Tahap awal penelitian analisis masalah dan spesifikasi kebutuhan difokuskan pada identifikasi masalah utama pada *chatbot* berbasis LLM untuk domain khusus, terutama risiko *hallucination* (jawaban tidak didukung sumber) dan *context* mismatch (konteks *retrieval* tidak tepat sehingga jawaban melenceng). Fenomena *hallucination* telah dipetakan sebagai isu penting dalam generasi bahasa alami dan menjadi motivasi kuat penggunaan *retrieval grounding* untuk meningkatkan keterlacakan dan ketepatan informasi [9]. Pada tahap ini juga ditetapkan ruang lingkup domain, karakteristik dokumen, serta kebutuhan sistem evaluasi *retrieval*. Penentuan metrik evaluasi dilakukan dengan memilih Recall@k dan MRR karena keduanya merupakan metrik standar dalam evaluasi information *retrieval* dan ranking, khususnya untuk menilai keberhasilan menemukan dokumen relevan dan kualitas urutan peringkat [12],[13].

Pengumpulan data domain khusus data penelitian berupa kumpulan dokumen domain khusus yang berfungsi sebagai *knowledge base* untuk *retrieval* (bukan data pelatihan LLM). Dokumen dihimpun dari sumber relevan seperti dokumentasi teknis/FAQ/panduan layanan agar skenario *retrieval* menyerupai penerapan RAG pada sistem nyata. Seluruh dokumen dikonversi menjadi teks untuk mendukung proses indexing dan pencarian, sebagaimana praktik umum pada sistem RAG yang memisahkan memori non-parametrik (koleksi dokumen) dari model generatif [8].

Tahap pra-pemrosesan bertujuan menyiapkan dokumen agar optimal untuk retrieval. Proses mencakup pembersihan elemen non-informatif, pemecahan dokumen menjadi unit teks (*chunking*) untuk meningkatkan granularitas pencarian, penambahan metadata (mis. halaman/sumber) untuk menjaga keterlacakan konteks. *Chunking* dan metadata penting karena kualitas *retrieval* dan ranking sangat dipengaruhi oleh bagaimana dokumen dipotong dan diindeks; kesalahan desain dapat memunculkan *ranking ambiguity*, yaitu dokumen benar muncul tetapi tidak berada pada posisi pertama [8], [12].

Perancangan Sistem *Retrieval-Augmented Generation* (RAG), arsitektur RAG dirancang dengan dua komponen utama *retriever* mengambil konteks paling relevan dari knowledge base, *generator* (LLM): menghasilkan jawaban menggunakan konteks hasil *retrieval*. Konsep RAG sebagai integrasi memori parametrik dan non-parametrik dirujuk dari formulasi RAG untuk tugas *knowledge-intensive* NLP. Dalam penelitian ini, fokus utama ditempatkan pada evaluasi *retrieval* (komponen *retriever*), karena kualitas konteks berpengaruh langsung terhadap akurasi jawaban dan potensi hallucination [8], [9].

Implementasi metode retrieval Tiga metode retrieval diimplementasikan dalam lingkungan eksperimen yang sama agar perbandingan adil *lexical retrieval* (TF-IDF/BM25), Metode leksikal menggunakan pembobotan term dan pencocokan kata sebagai *baseline* klasik IR [5]. BM25 sebagai varian probabilistik yang sangat umum digunakan juga menjadi rujukan pendekatan leksikal modern [12]. *Semantic* Retrieval berbasis *Vector*

Embedding Pendekatan *dense retrieval* memetakan *query* dan dokumen ke ruang vektor, lalu melakukan pencarian berdasarkan kemiripan *embedding*. Pendekatan ini ditunjukkan efektif pada open-domain QA melalui *Dense Passage Retrieval* (DPR) [10]. Implementasi semantic retrieval umumnya menggunakan mesin pencarian kemiripan vektor berskala besar seperti FAISS untuk efisiensi *nearest neighbor search* [15].

Hybrid Retrieval (*Lexical + Semantic*) dirancang untuk menggabungkan keunggulan pencocokan terminologi (leksikal) dan pemahaman makna (semantik). Penggabungan ranking dapat dilakukan melalui teknik fusi seperti *Reciprocal Rank Fusion* (RRF) yang terbukti meningkatkan performa penggabungan beberapa sistem IR [16]. Contoh penggunaan kombinasi sinyal *sparse/dense* juga ditunjukkan pada sistem IR semantik yang menggabungkan representasi TF-IDF/BM25 dengan *encoder* semantik [11]. Pada tahap ini, sistem menghasilkan *ranked* list dokumen untuk setiap query uji tanpa intervensi manual.

Evaluasi Metode Retrieval (Inti Penelitian) Evaluasi dilakukan menggunakan metrik kuantitatif standar IR: Recall@1, Recall@3, Recall@5, dan MRR [12],[13]. Recall@k mengukur apakah dokumen benar muncul dalam top-k hasil retrieval. MRR menilai seberapa cepat dokumen benar muncul di peringkat atas (sangat relevan untuk sistem RAG karena top-1 ranking biasanya menjadi konteks utama). Selain evaluasi kuantitatif, dilakukan error analysis untuk mengidentifikasi pola kegagalan, khususnya kasus ketika dokumen benar muncul di peringkat 2–3 (indikasi ranking *ambiguity*, bukan *retrieval failure*). Fokus evaluasi ini selaras dengan tujuan penelitian yang menilai kualitas retrieval dan ranking sebagai fondasi reliabilitas RAG [8] [9].

Integrasi Model RAG ke Website (Validasi Sistem) dengan performa terbaik dari tahap evaluasi diintegrasikan ke prototipe *chatbot* berbasis web sebagai validasi end-to-end. Tujuan integrasi adalah memastikan *pipeline* RAG (retrieval → konteks → generasi) dapat berjalan stabil pada skenario penggunaan nyata. Tahap validasi implementatif seperti ini telah diterapkan dalam studi RAG berbasis web/layanan informasi pada publikasi nasional (SINTA) yang mengadopsi RAG untuk *chatbot* domain tertentu [15][16].

3. HASIL DAN ANALISIS

Berdasarkan Tahap 1 metodologi penelitian, analisis masalah menunjukkan bahwa *chatbot* berbasis model bahasa besar pada domain khusus menghadapi dua permasalahan utama, yaitu *hallucination* dan ketidaktepatan pemilihan konteks (*context mismatch*). Permasalahan ini terutama muncul ketika model menghasilkan jawaban tanpa dukungan dokumen domain yang tepat. Hasil analisis juga menunjukkan bahwa domain khusus yang diteliti memiliki karakteristik dokumen teknis dengan terminologi eksplisit dan struktur informasi yang relatif tetap. Oleh karena itu, sistem *chatbot* yang dikembangkan membutuhkan mekanisme *retrieval* yang mampu menemukan dokumen yang benar, dan memprioritaskan dokumen tersebut pada peringkat teratas. Berdasarkan kebutuhan tersebut, metrik Recall@k dan *Mean Reciprocal Rank* (MRR) ditetapkan sebagai indikator evaluasi utama karena mampu mengukur keberhasilan penemuan dokumen dan kualitas urutan peringkat konteks.

Hasil pengumpulan data domain khusus pada tahap ini diperoleh kumpulan dokumen domain khusus yang digunakan sebagai knowledge base retrieval. Dokumen yang terkumpul mencakup dokumentasi teknis, panduan penggunaan, dan halaman pendukung sistem yang relevan dengan domain penelitian. Seluruh dokumen berhasil dikonversi ke format teks tanpa kehilangan informasi penting. Hasil tahap ini memastikan bahwa data yang digunakan bersifat representatif terhadap kebutuhan informasi pengguna dalam skenario chatbot domain khusus.

Hasil pra-pemrosesan data menunjukkan bahwa proses pra-pemrosesan berhasil menghasilkan kumpulan dokumen terstruktur yang siap diindeks. Dokumen diproses melalui pembersihan teks, pemecahan menjadi unit teks (*chunking*), serta penambahan metadata halaman. Observasi pada tahap ini menunjukkan bahwa ukuran chunk dan keberadaan metadata memiliki pengaruh langsung terhadap granularitas retrieval. Dokumen yang terlalu besar cenderung menghasilkan konteks yang umum, sedangkan dokumen yang terlalu kecil meningkatkan risiko fragmentasi konteks. Temuan ini menjadi dasar untuk memahami munculnya fenomena ranking *ambiguity* pada tahap evaluasi.

Hasil perancangan sistem *Retrieval-Augmented Generation* menghasilkan rancangan arsitektur sistem RAG yang memisahkan komponen *retriever* dan generator (LLM). Sistem dirancang agar satu komponen retrieval dapat diuji menggunakan beberapa metode tanpa mengubah komponen generatif. Tiga pendekatan retrieval berhasil dirancang yaitu TF-IDF sebagai pendekatan leksikal, *Semantic retrieval* berbasis *vector embedding*, dan *Hybrid retrieval* yang menggabungkan sinyal leksikal dan semantik. Rancangan ini memungkinkan evaluasi yang adil terhadap setiap metode retrieval karena seluruh metode menggunakan *dataset, query uji*, dan skema evaluasi yang sama.

Hasil implementasi metode retrieval, ketiga metode retrieval berhasil diimplementasikan dalam lingkungan eksperimen yang sama. Setiap *query* uji menghasilkan daftar peringkat dokumen (*ranked list*) untuk masing-masing metode. Hasil implementasi menunjukkan bahwa seluruh metode dapat berfungsi secara stabil dan

konsisten, serta menghasilkan keluaran retrieval yang siap dievaluasi secara kuantitatif. Tidak ditemukan kegagalan sistem atau missing output selama proses pengujian.

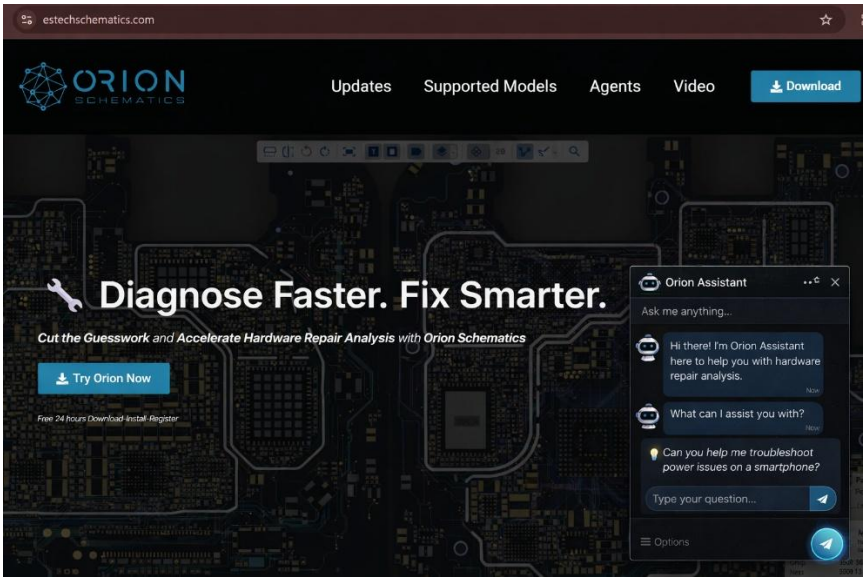
Tabel 1. Hasil Evaluasi Kinerja Metode Retrieval

Metode Retrieval	*Recall@1*	*Recall@3*	*Recall@5*	*MRR*
TF-IDF	0.625	0.85	0.8625	0.7233
Semantic	0.35	0.4875	0.625	0.4408
Hybrid	0.4	0.7125	0.825	0.5542

Hasil pada Tabel 1 menunjukkan bahwa TF-IDF memperoleh performa tertinggi pada Recall@1 dan MRR, menandakan keunggulan dalam menempatkan dokumen relevan pada peringkat teratas. *Semantic retrieval* menunjukkan performa terendah pada seluruh metrik evaluasi, sementara *hybrid retrieval* berada di antara kedua metode tersebut. *Hybrid retrieval* secara signifikan meningkatkan Recall@3 dan Recall@5 dibandingkan *semantic retrieval*, yang menunjukkan peningkatan cakupan dokumen relevan. Namun, nilai Recall@1 dan MRR *hybrid* masih berada di bawah TF-IDF, mengindikasikan bahwa peningkatan cakupan belum sepenuhnya diikuti oleh peningkatan kualitas peringkat teratas.

Hasil analisis kesalahan dan *ranking ambiguity* Berdasarkan hasil evaluasi retrieval untuk contoh pertanyaan "Siapa yang cocok menggunakan *Orion Schematics*?", ketiga pendekatan menunjukkan kemampuan menemukan dokumen yang relevan, namun dengan karakteristik yang berbeda. TF-IDF berhasil menempatkan halaman FAQ utama (PAGE 1) pada peringkat teratas, yang secara eksplisit menjelaskan target pengguna *Orion Schematics* seperti teknisi, *engineer*, dan profesional perbaikan perangkat elektronik, tetapi juga mengangkat dokumen administratif yang kurang sesuai dengan intent pertanyaan. *Semantic retrieval* menunjukkan pemahaman makna yang lebih luas dengan turut mengidentifikasi dokumen terkait segmen pengguna enterprise dan institusi, namun mengalami *semantic drift* karena beberapa dokumen operasional dan persyaratan sistem ikut dianggap relevan. Pendekatan *hybrid* (RRF) mampu mengombinasikan keunggulan kedua metode dengan mempertahankan dokumen deskriptif utama sekaligus memperluas cakupan segmen pengguna, tetapi perbedaan skor antar dokumen yang sangat kecil mengindikasikan terjadinya ranking ambiguity. Temuan ini menunjukkan bahwa untuk pertanyaan kategorikal-deskriptif, tantangan utama dalam sistem RAG bukan pada kegagalan menemukan informasi, melainkan pada pemilihan dan pemrioritasan konteks yang paling representatif untuk mendukung generasi jawaban yang ringkas dan tepat sasaran.

hasil integrasi model RAG ke website, metode retrieval dengan performa terbaik diintegrasikan ke dalam prototipe *chatbot* berbasis web sebagai validasi sistem. Hasil integrasi menunjukkan bahwa sistem RAG dapat berjalan secara end-to-end dan mampu memberikan respons berbasis konteks hasil retrieval. Untuk disain menunya sebagaimana gambar 2. Disain tersebut menu *chatbot* ditambahkan di menu utama web Perusahaan. Untuk nama *chatbot*nya adalah *Orion Asistant*.



Gambar 2 Disain menu chatbot

4. KESIMPULAN

Penelitian ini mengevaluasi efektivitas metode retrieval dalam implementasi chatbot domain khusus berbasis Retrieval-Augmented Generation (RAG) untuk meningkatkan akurasi dan relevansi informasi serta

mengurangi risiko hallucination. Hasil evaluasi menunjukkan bahwa pendekatan TF-IDF masih unggul dalam menempatkan dokumen relevan pada peringkat teratas pada domain dengan terminologi teknis yang eksplisit, sementara semantic retrieval murni mengalami keterbatasan akibat semantic drift, dan pendekatan hybrid mampu meningkatkan cakupan dokumen relevan pada peringkat menengah namun belum optimal dalam memprioritaskan konteks paling representatif. Analisis lebih lanjut mengungkapkan bahwa permasalahan utama sistem RAG pada domain khusus bukan terletak pada kegagalan menemukan informasi, melainkan pada ranking ambiguity dalam menentukan prioritas konteks untuk mendukung generasi jawaban yang akurat. Dengan demikian, kontribusi utama penelitian ini adalah menunjukkan bahwa peningkatan reliabilitas chatbot berbasis RAG pada domain khusus lebih ditentukan oleh kualitas penentuan peringkat konteks daripada kemampuan menemukan dokumen semata, sekaligus menegaskan pentingnya evaluasi metode retrieval sebagai fondasi pengembangan chatbot RAG yang andal dan kontekstual, serta membuka peluang pengembangan lanjutan melalui optimalisasi mekanisme re-ranking dan strategi retrieval adaptif berbasis jenis pertanyaan.

REFERENSI

[1] Zhou et al. "GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation" Frontiers in medicine (2024).

[2] Church et al. "Emerging trends: a gentle introduction to RAG" Natural language engineering (2024)

[3] Ji et al. "Survey of Hallucination in Natural Language Generation" ACM Computing Surveys (2023)

[4] Li et al. "VaxBot-HPV: a GPT-based chatbot for answering HPV vaccine-related questions" Jamia Open (2024).

[5] Rukhiran and Netinant "Automated information retrieval and services of graduate school using chatbot system" International Journal of Electrical and Computer Engineering (2022).

[6] Son et al. "Development and Evaluation of a Retrieval-Augmented Generation-Based Electronic Medical Record Chatbot System" Healthcare Informatics Research (2025).

[7] Nishisako et al. "Reducing Hallucinations and Trade-Offs in Responses in Generative AI Chatbots for Cancer Information: Development and Evaluation Study" JMIR Cancer (2025).

[8] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020..

[9] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, 2023.

[10] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020..

[11] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, 2023.

[10] V. Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering,” EMNLP, 2020.

[11] A. Esteva et al., “CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization,” arXiv, 2020.

[12] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2009 (online ed.).

[13] T. Sakai, “Test Collection Based Evaluation of Information Retrieval Systems,” Foundations and Trends in Information Retrieval, 2010 (overview metrik evaluasi IR).

[14] S. Robertson, “The Probabilistic Relevance Framework: BM25 and Beyond,” 2009.

[15] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” arXiv, 2017.

[16] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, “Reciprocal Rank Fusion outperforms Condorcet and individual rank learning methods,” SIGIR, 2009.

[17] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu, “Okapi at TREC-3,” 1994.

[18] G. H. Setiawan and I. M. B. Adnyana, “Improving Helpdesk Chatbot Performance with TF-IDF and Cosine Similarity Models,” JAIC, vol. 7, no. 2, pp. 252–257, 2023.

[19] G. Samudra, A. Turmudi Zy, and Ermanto, "Implementation of Retrieval Augmented Generation (RAG) in the Design of Digestive Health Chatbot," JSAI, vol. 8, no. 1, pp. 181–188, 2025.

[20] D. Firdaus, I. Sumardi, and Y. Kulsum, "Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7B for Indonesian Medical Herb," JISKA, vol. 9, no. 3, pp. 230–243, 2024.

[14] G. D. Albert and A. Voutama, "Pengembangan Chatbot Berbasis PDF Menggunakan Local Retrieval-Augmented Generation (RAG) dan Ollama," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 2, 2025.

[15] L. Anindyati, "Analisis dan Perancangan Aplikasi Chatbot Menggunakan Framework Rasa …," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 10, no. 2, pp. 291–300, 2023.

[16] P. A. Faranisa et al., "Klasifikasi Layanan Informasi … Retrieval-based Chatbot Menggunakan Naïve Bayes dan TF-IDF," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK), 2024.