Implementasi Chatbot Berbasis *Large Language Model* dengan *Retrieval-Augmented Generation* untuk Pengetahuan Perusahaan

Slamet Meisa Putra [a,1,*], Fitriasih [b,2]

[a b] Teknik Informatika, Politeknik Purbaya Tegal, Indonesia
[1] pakrt556@gmail.com*; [2] pipit.qolbu@gmail.com;
* Penulis Korespondensi

ABSTRAK

PT Java Abadi Gemilang merupakan perusahaan jasa teknologi informasi yang bekerja sama dengan HP Service Center. Sistem saluran informasi internal yang digunakan masih bersifat manual, sehingga menyebabkan rendahnya efisiensi, keterlambatan penyampaian dan kesulitan pencarian informasi oleh karyawan. Penelitian ini bertujuan mengatasi masalah tersebut dengan membuat sebuah sistem Chatbot berbasis *Large Language Model* dengan pendekatan *Retrieval-Augmented Generation*, yang memiliki kemampuan untuk memahami permintaan pengguna dan memberikan respons informasi secara cepat dan akurat. Penelitian ini menggunakan pendekatan penelitian terapan dengan tahapan yang meliputi Studi literatur, Pengumpulan dan Pra-pemrosesan data informasi internal perusahaan, Merancang dan Implementasi Sistem Chatbot, Pengujian sistem dan evaluasi. Hasil penelitian menunjukkan bahwa Chatbot dapat memahami permintaan pengguna dan memberikan respons informasi secara cepat dan relevan berdasarkan dokumen internal. Implementasi sistem ini meningkatkan efisiensi pencarian informasi dan kualitas layanan internal perusahaan. Penelitian ini menyimpulkan bahwa Chatbot berbasis *Large Language Model* dengan *Retrieval-Augmented Generation* efektif digunakan sebagai saluran informasi perusahaan.

Riwayat Artikel
Diterima 10 Januari 2024
Diperbaiki 15 Februari 2024
Diterbitkan 20 Maret 2024

Kata Kunci
Large Language Model
Retrieval-Augmented Generation
Chatbot
Sistem Informasi Internal

1. Pendahuluan

Dalam beberapa tahun terakhir, perkembangan kecerdasan buatan (Artificial Intelligence) di bidang teknologi informasi menunjukkan kemajuan yang signifikan, salah satunya melalui penerapan chatbot. Sejumlah studi mengungkapkan bahwa penggunaan chatbot mampu mendukung peningkatan efisiensi layanan melalui otomatisasi penyediaan informasi, percepatan waktu respons, pengurangan beban kerja manual serta efisiensi biaya operasional [1], [2], [3], [4]. Meskipun demikian, berdasarkan hasil observasi dan pengalaman penulis selama mengikuti program Merdeka Belajar Kampus Merdeka di Service Center PT. Java Abadi Gemilang, pemanfaatan chatbot sebagai media informasi internal perusahaan hingga saat ini belum diterapkan. Padahal, teknologi tersebut berpotensi meningkatkan efektivitas layanan internal.

Pemanfaatan chatbot berbasis Large Language Model (LLM) memungkinkan pemrosesan bahasa alami yang lebih adaptif terhadap konteks komunikasi pengguna. Dengan menganalisis masukan dalam bentuk teks maupun suara, sistem ini mampu mengenali informasi penting dan menyesuaikan respons, sehingga pola interaksi yang terbentuk menjadi lebih mendekati percakapan antar manusia [5], [6]. Akan tetapi pemanfaatan Large Language Model masih memiliki kelemahan, salah satunya kecenderungan menghasilkan jawaban yang terdengar meyakinkan tetapi tidak sesuai fakta, sebuah fenomena yang sering disebut halusinasi [7], [8]. Seperti dijelaskan oleh Magdalena Wysocka dan rekan-rekan (2024), “model yang lebih mutakhir memang kian lancar dalam menyusun kalimat, tetapi ketepatan faktualnya masih belum stabil, dan bias terhadap entitas yang sering muncul dalam data pelatihan masih tampak jelas” [9].

e-ISSN: 2809-3399, p-ISSN: 2338-5197 Vol. 14., No. 1, Februari 2026, pp. 53-62

Untuk meminimalkan masalah tersebut, pendekatan Retrieval-Augmented Generation (RAG) menjadi solusi yang semakin banyak dibicarakan. Murtiyoso dan kolega (2025) menjelaskan bahwa "Retrieval-Augmented Generation bekerja dengan menggabungkan mekanisme pencarian informasi eksternal dan kemampuan generatif Large Language Model, sehingga model dapat mengakses sumber pengetahuan tambahan di luar data pelatihannya. Dengan bantuan informasi dari luar, kualitas dan ketelitian jawaban yang dihasilkan meningkat, terutama pada bidang yang memerlukan pengetahuan rinci dan selalu berubah" [10]. Secara teknis, Retrieval-Augmented Generation bekerja dengan memanfaatkan dokumen eksternal, seperti berkas PDF, yang diproses menjadi representasi vektor melalui teknik embedding. Vektor-vektor ini kemudian disimpan dalam basis data dan digunakan sebagai sumber referensi tambahan yang terintegrasi dengan sistem chatbot, sehingga model dapat menghasilkan jawaban berdasarkan konteks dan data yang relevan [11]. Dengan menggunakan pendekatan ini, chatbot dapat memberikan jawaban yang tidak hanya natural tetapi berdasarkan bukti faktual dari data internal perusahaan seperti SOP, kebijakan dan aturan layanan, FAQ internal.

Berbagai penelitian telah menerapkan Retrieval-Augmented Generation (RAG) dalam pengembangan chatbot. Penelitian oleh Rizki Dharma Raya dan F. Lia Dwi Cahyanti menunjukkan bahwa integrasi RAG dengan berbagai model seperti Llama, Gemma, dan Mistral mampu menghasilkan respons yang akurat dan relevan secara kontekstual [12]. Sementara itu, studi Gerald Dustin Albert dan Apriade Voutama melaporkan bahwa chatbot berbasis PDF dengan pendekatan local RAG mampu mencapai ROUGE-L sebesar 0,85 dengan waktu respons 5–10 detik [13]. Temuan tersebut menegaskan efektivitas RAG dalam meningkatkan kesesuaian jawaban terhadap dokumen referensi.

Di sisi lain, penelitian Muhammad Syarifudin et al. lebih menitikberatkan pada aspek kualitas layanan dan dampaknya terhadap perilaku pengguna, menunjukkan bahwa kualitas layanan AI-chatbot berpengaruh signifikan terhadap kepercayaan dan keterlibatan pengguna [14]. Meskipun demikian, sebagian besar penelitian tersebut berfokus pada implementasi umum atau konteks layanan publik dan e-commerce, serta belum secara spesifik mengkaji karakteristik basis pengetahuan internal perusahaan yang bersifat prosedural dan dinamis. Evaluasi yang dilakukan juga umumnya tidak membandingkan secara eksplisit performa sistem dengan dan tanpa mekanisme retrieval dalam konteks operasional perusahaan. Oleh karena itu, penelitian ini berupaya mengisi celah tersebut dengan mengimplementasikan dan mengevaluasi RAG secara komparatif pada sistem informasi internal berbasis dokumen SOP perusahaan.

2. Metode

Penelitian merupakan jenis penelitian terapan yang dimana memiliki beberapa tahapan yaitu Studi literatur, Pengumpulan dan Pra-pemrosesan data informasi internal perusahaan, Merancang dan Implementasi Sistem Chatbot, Pengujian sistem dan evaluasi, seperti pada gambar 1 di bawah ini:

2.1. Studi Literatur

Studi Literatur merupakan tahap awal dalam penelitian ini yang dilakukan melalui proses pengumpulan dan analisis sumber-sumber ilmiah dari berbagai basis data, seperti Google Scholar, Garuda, Scopus, dan SINTA. Tahap ini bertujuan untuk memperoleh pemahaman komprehensif yang mendukung perumusan metodologi penelitian, khususnya terkait implementasi chatbot berbasis Large Language Model (LLM) dengan pendekatan Retrieval-Augmented Generation (RAG).

2.2. Pengumpulan dan Pra-Pemrosesan Data Internal

Tahap ini meliputi pengumpulan data internal perusahaan berupa dokumen SOP, kebijakan layanan, serta FAQ internal dalam format PDF. Seluruh dokumen yang diperoleh akan melalui serangkaian proses pra-pemrosesan data yang terdiri atas beberapa tahapan seperti berikut:

2.2.1.Data Cleaning

Tahap data cleaning dilakukan untuk menghilangkan noise pada dokumen, seperti duplikasi informasi, karakter tidak relevan, header dan footer otomatis, serta ketidakkonsistenan format teks hasil ekstraksi PDF, sehingga diperoleh data teks yang bersih dan siap diproses lebih lanjut.

2.2.2. Text Splitting (Chunking)

Tahap text splitting dilakukan dengan membagi dokumen yang telah dibersihkan menjadi potongan teks (chunk) berukuran sekitar 1000 token per bagian. Ukuran chunk yang relatif besar dipilih untuk mempertahankan konteks prosedural yang utuh, khususnya pada dokumen SOP yang memiliki alur langkah berurutan. Pendekatan ini bertujuan mengurangi fragmentasi informasi yang dapat menurunkan relevansi hasil retrieval pada sistem Retrieval-Augmented Generation.

2.3. Merancang dan Implementasi Sistem

Tahap ini berfokus pada perancangan arsitektur chatbot berbasis platform n8n yang memanfaatkan Large Language Model (LLM) dengan pendekatan Retrieval-Augmented Generation (RAG) sebagai layanan informasi internal. Interaksi dimulai ketika pengguna mengirim pesan melalui Telegram yang kemudian diproses oleh AI Agent terintegrasi dengan Google Gemini sebagai model generatif. Dokumen internal yang telah dibersihkan dan dipotong menjadi beberapa bagian diubah ke dalam bentuk embedding untuk merepresentasikan makna teks secara numerik. Embedding tersebut disimpan pada Supabase Vector Store guna memungkinkan pencarian informasi berbasis kesamaan makna (semantic similarity). Ketika pengguna mengajukan pertanyaan, sistem mengubah query menjadi embedding dan melakukan pencarian menggunakan cosine similarity untuk memperoleh tiga dokumen paling relevan (top-k = 3). Pemilihan jumlah tersebut bertujuan menjaga keseimbangan antara kelengkapan konteks dan meminimalkan informasi yang tidak relevan. Hasil pencarian kemudian dimasukkan ke dalam prompt sebelum diproses oleh model LLM yang dikonfigurasi dengan temperatur rendah untuk meningkatkan konsistensi jawaban serta mengurangi potensi halusinasi. Respons akhir selanjutnya dikirim kembali kepada pengguna melalui Telegram.

2.4. Pengujian dan Evaluasi Sistem

Tahap pengujian dilakukan untuk menilai tingkat ketepatan respons chatbot serta membandingkan kinerjanya pada skenario dengan dan tanpa penerapan Retrieval-Augmented Generation (RAG). Jumlah pertanyaan uji diperluas dari 15 menjadi 30 untuk meningkatkan variasi konteks dan validitas evaluasi. Evaluasi dilakukan dalam dua tahap. Tahap pertama menggunakan metode Human Semantic Evaluation, dengan membandingkan jawaban chatbot terhadap dokumen SOP perusahaan sebagai referensi utama. Penilaian didasarkan pada tiga indikator, yaitu relevansi, kelengkapan informasi, dan akurasi fakta. Respons kemudian diklasifikasikan ke dalam kategori “Sesuai”, “Sebagian Sesuai”, atau “Tidak Sesuai”. Metode Human Semantic Evaluation dipilih karena lebih sesuai untuk sistem Question Answering berbasis dokumen SOP yang bersifat prosedural, dimana ketepatan makna, konteks operasional, dan kelengkapan informasi lebih relevan dinilai secara semantik oleh evaluator manusia dibandingkan metrik otomatis berbasis kesamaan teks. Tahap kedua menganalisis performa sistem berdasarkan mekanisme pembangkitan jawaban (RAG dan non-RAG). Tingkat akurasi dihitung dari proporsi jawaban yang dikategorikan “Sesuai” terhadap total pertanyaan menggunakan rumus:

$$Akurasi\ (\%) = (Jumlah\ jawaban\ \text{“}Sesuai\text{”}\ /\ Total\ pertanyaan) \times 100\%$$

Nilai akurasi tersebut digunakan untuk membandingkan efektivitas kedua pendekatan dalam menghasilkan respons yang sesuai dengan dokumen internal. [15].

Jurnal Sarjana Teknik Informatika
Vol. 14., No. 1, Februari 2026, pp. 53-62

3. Hasil dan Pembahasan

Hasil dan pembahasan menjelaskan implementasi chatbot serta menganalisis kelayakan sistem yang menerapkan metode *Retrieval-Augmented Generation* (RAG). Basis pengetahuan berbasis dokumen SOP diproses melalui embedding dan disimpan pada vector database, sehingga pembaruan informasi dapat dilakukan secara dinamis tanpa pelatihan ulang model. Mekanisme ini memastikan proses retrieval selalu mengacu pada dokumen terbaru sehingga konteks jawaban tetap relevan dengan prosedur operasional perusahaan.

Evaluasi dilakukan untuk menilai kemampuan pendekatan RAG dalam menghasilkan respons yang akurat dan kontekstual menggunakan 30 pertanyaan uji yang diklasifikasikan menjadi pertanyaan eksplisit, implisit, dan di luar konteks basis pengetahuan. Penggunaan chunk berukuran besar pada proses retrieval membantu mempertahankan konteks prosedural dokumen internal sehingga respons menjadi lebih konsisten pada pertanyaan eksplisit, tetap relevan pada pertanyaan implisit yang memerlukan inferensi kontekstual, serta mampu menolak pertanyaan yang tidak berkaitan dengan dokumen SOP.

Selain aspek akurasi, implementasi chatbot berbasis RAG juga memiliki implikasi terhadap keamanan dan privasi informasi internal. Meskipun dokumen SOP disimpan dalam bentuk embedding pada vector database, representasi tersebut tetap mempertahankan makna semantik sehingga berpotensi diekstraksi melalui interaksi berulang dengan sistem. Oleh karena itu, chatbot dirancang sebagai sistem internal yang hanya dapat diakses dalam lingkungan operasional perusahaan dengan pembatasan penggunaan untuk kebutuhan informasi prosedural karyawan. Pendekatan ini bertujuan menjaga keseimbangan antara kemudahan akses informasi dan perlindungan basis pengetahuan organisasi.

Gambar 6 Contoh Teks Respon Bot

Gambar 6 menunjukkan interaksi awal pengguna dengan chatbot ketika memulai percakapan. Pada contoh tersebut, pengguna memasukkan sapaan “selamat pagi” dan sistem merespons dengan sapaan balasan yang relevan, yaitu “selamat pagi juga! Ada yang bisa saya bantu?”. Respons ini menunjukkan bahwa chatbot mampu mengenali intent dasar berupa sapaan dan memberikan tanggapan yang sesuai dengan konteks percakapan.

Gambar 7 Contoh Teks Respon Bot

Gambar 7 menunjukkan contoh interaksi ketika pengguna menanyakan media pengajuan klaim service atau garansi. Pengguna memasukkan pertanyaan mengenai saluran yang dapat digunakan untuk mengajukan klaim, kemudian chatbot merespons bahwa klaim dapat dilakukan melalui nomor Service Center seperti WhatsApp, SMS, dan telepon, serta media lain yang akan diteruskan kepada product specialist. Respons tersebut sesuai dengan prosedur yang tercantum dalam SOP perusahaan.

Table 1: Evaluasi Kualitas Respons Chatbot

NO	KATEGORI	PERTANYAAN	RESPON CHATBOT	SUMBER SOP	PENILAIAN SEMANTIK
1.	Eksplisit	Berapa batas waktu maksimal untuk perbaikan ringan di Service Center dan perbaikan teknisi di Warehouse, serta apa tujuan penetapan waktu tersebut?	Perbaikan ringan di Service Center harus diselesaikan dalam waktu maksimal 2 jam pada jam kerja. Untuk perbaikan di Warehouse oleh teknisi, batas waktunya adalah maksimal 2 hari kerja setelah laptop diterima. Tujuan penetapan batas waktu ini adalah untuk memastikan efisiensi dan kecepatan dalam proses perbaikan, sehingga kepuasan pelanggan dapat terjaga dengan baik.	Ya	Sesuai
2.	Eksplisit	Melalui media apa saja customer dapat mengajukan klaim service atau garansi?	Untuk mengajukan klaim service atau garansi, customer dapat menghubungi kami melalui: Telepon, Email, Formulir kontak di website kami	Tidak	Sebagian Sesuai

NO	KATEGORI	PERTANYAAN	RESPON CHATBOT	SUMBER SOP	PENILAIAN SEMANTIK
3.	Implisit	Apa tujuan dilakukan pemeriksaan menyeluruh meskipun hanya satu komponen yang dilaporkan rusak?	Tujuan dilakukannya pemeriksaan menyeluruh meskipun hanya satu komponen yang dilaporkan rusak adalah untuk mengidentifikasi potensi masalah lain yang mungkin belum terlihat atau belum dilaporkan.	Tidak	Tidak Sesuai
4.	Implisit	Apa manfaat melakukan general check seluruh komponen laptop setelah perbaikan utama?	General check setelah perbaikan utama penting untuk memastikan semua komponen berfungsi optimal dan terintegrasi dengan baik.	Tidak	Tidak Sesuai
5.	Diluar Konteks	Berapa rata-rata biaya perbaikan kerusakan berat laptop di Service Center?	Maaf, saya tidak dapat menemukan informasi mengenai rata-rata biaya perbaikan kerusakan berat laptop di database..	Tidak	Sesuai
6.	Diluar Konteks	Merek laptop apa yang paling sering mengalami klaim garansi?	Maaf, data yang Anda minta tidak tersedia. Informasi mengenai merek laptop yang paling sering mengalami klaim garansi tidak terdapat dalam database.	Ya	Sesuai

Table 2: Evaluasi Performa RAG

Kategori	Pertanyaan	RAG	Non RAG	Presentase RAG	Presentase Non RAG
Eksplisit	Mengapa perbaikan ringan tidak perlu dikirim ke Warehouse?	✓		90%	10%
	Berapa batas waktu maksimal untuk perbaikan ringan di Service Center dan perbaikan teknisi di Warehouse, serta apa tujuan penetapan waktu tersebut ?	✓			
	Jika saya mengalami masalah dengan lupa kode kunci, termasuk kategori perbaikan apa dan berapa lama perbaikan tersebut ?	✓			
	Jelaskan alur pengiriman laptop dari service center ke Warehouse hingga kembali ke customer ?		✓		
	Melalui media apa saja customer dapat mengajukan klaim service atau garansi ?	✓			

Kategori	Pertanyaan	RAG	Non RAG	Presentase RAG	Presentase Non RAG
	Bagian mana yang melakukan proses pengepakan ulang sebelum perangkat dikirim kembali ke customer?	✓			
	Apa yang harus dilakukan petugas sebelum perangkat diterima secara resmi dari customer di Service Center?	✓			
	Pada tahap mana perangkat dilakukan pengecekan ulang sebelum dikirim kembali kepada customer?	✓			
	Siapa pihak yang menangani perangkat ketika perbaikan tidak dapat dilakukan di Service Center?	✓			
	Apa tindakan yang dilakukan setelah proses perbaikan perangkat selesai sebelum dilakukan pengiriman?	✓			
Implisit	Mengapa customer harus selalu mendapat update setiap tahap perbaikan dan pengiriman ?	✓		80 %	20%
	Mengapa dokumentasi foto atau video kerusakan diperlukan dalam proses service ?	✓			
	Apa risiko yang mungkin terjadi jika verifikasi kartu garansi tidak dilakukan dengan database e-commerce ?	✓			
	Apa manfaat melakukan general check seluruh komponen laptop setelah perbaikan utama ?		✓		
	Apa dampak jika update penerimaan laptop tidak diberikan kepada customer?	✓			
	Mengapa proses pengecekan kelengkapan perangkat perlu dilakukan dua kali selama proses service?	✓			
	Bagaimana prosedur dokumentasi membantu mengurangi potensi sengketa antara customer dan Service Center?	✓			
	Mengapa perangkat perlu dibersihkan dan dirapikan sebelum dikirim kembali ke customer?	✓			
	Apa tujuan dilakukan pemeriksaan menyeluruh meskipun hanya satu komponen yang dilaporkan rusak?		✓		
	Mengapa koordinasi antara Service Center dan Warehouse penting dalam menjaga kualitas layanan?	✓			
Diluar konteks	Berapa rata-rata biaya perbaikan kerusakan berat laptop di Service Center ?		✓	80%	20%
	Merek laptop apa yang paling sering mengalami klaim garansi ?	✓			
	Berapa jumlah Teknisi yang bekerja di warehouse setiap hari ?		✓		
	Berapa lama rata-rata pengiriman ekspedisi dari warehouse ke rumah customer ?	✓			
	Apakah Service Center memberikan garansi ulang setelah perbaikan selesai ?	✓			
	Berapa tingkat kepuasan pelanggan terhadap layanan Service Center dalam satu tahun terakhir?	✓			
	Apakah Service Center menyediakan layanan perbaikan di lokasi customer (home service)?	✓			

e-ISSN: 2809-3399, p-ISSN: 2338-5197 Vol. 14., No. 1, Februari 2026, pp. 53-62

Kategori	Pertanyaan	RAG	Non RAG	Presentase RAG	Presentase Non RAG
	Apakah perusahaan menyediakan layanan prioritas berbayar untuk mempercepat proses service?	✓			
	Apakah teknisi Service Center menggunakan kecerdasan buatan untuk mendiagnosis kerusakan laptop secara otomatis?	✓			
	Apakah Service Center menyediakan layanan upgrade spesifikasi laptop sambil menunggu proses garansi?	✓			

Tabel evaluasi disusun untuk mengukur perbedaan tingkat akurasi respons chatbot berdasarkan kategori konteks, yaitu eksplisit, implisit, dan di luar konteks, pada skenario dengan penerapan *Retrieval-Augmented Generation* (RAG) maupun tanpa RAG. Respons diklasifikasikan sebagai RAG apabila sistem menghasilkan jawaban melalui proses retrieval dokumen, sedangkan Non-RAG merujuk pada jawaban yang dihasilkan tanpa proses retrieval.

Akurasi dihitung berdasarkan jumlah jawaban yang dikategorikan "Sesuai" pada hasil evaluasi semantik dibandingkan dengan total pertanyaan pada masing-masing kategori. Pada kategori eksplisit terdapat sepuluh pertanyaan. Dari jumlah tersebut, sembilan respons berbasis RAG dinilai "Sesuai", sehingga menghasilkan akurasi 9/10 × 100% = 90%, sedangkan satu respons non-RAG dinilai "Sesuai" dengan akurasi 10%. Pada kategori implisit, delapan respons berbasis RAG dinilai "Sesuai" (80%), sementara dua respons non-RAG dinilai "Sesuai" (20%). Pada kategori di luar konteks, delapan respons RAG dinilai sesuai (80%) dan dua respons non-RAG dinilai sesuai (20%). Secara keseluruhan, jumlah jawaban benar pada skenario RAG adalah 25 dari 30 pertanyaan, menghasilkan akurasi 83,3%. Sementara itu, skenario Non-RAG menghasilkan lima jawaban benar dari 30 pertanyaan, dengan akurasi 16,7%.

Peningkatan akurasi pada skenario RAG menunjukkan bahwa integrasi mekanisme retrieval berperan signifikan dalam meningkatkan kesesuaian jawaban terhadap dokumen SOP. Performa yang tinggi pada kategori eksplisit mengindikasikan efektivitas retrieval dalam menangani informasi prosedural terstruktur. Perbedaan performa pada kategori implisit menunjukkan bahwa meskipun retrieval membantu menyediakan konteks, sistem masih memiliki keterbatasan dalam melakukan penggabungan informasi lintas potongan teks.

4. Kesimpulan

Penelitian ini berhasil mengimplementasikan sistem chatbot berbasis Large Language Model (LLM) dengan pendekatan *Retrieval-Augmented Generation* (RAG) sebagai saluran informasi internal perusahaan. Hasil evaluasi terhadap 30 pertanyaan uji menunjukkan bahwa 83,3% respons berbasis RAG dinilai relevan atau akurat, dibandingkan 16,7% pada skenario tanpa RAG. Temuan ini menunjukkan bahwa integrasi mekanisme retrieval secara signifikan meningkatkan ketepatan konteks jawaban chatbot. Penerapan RAG terbukti efektif dalam menangani informasi prosedural yang terstruktur, seperti dokumen SOP, kebijakan layanan, dan FAQ internal, sehingga mampu meningkatkan efisiensi pencarian informasi bagi karyawan.

Meskipun demikian, performa sistem masih bergantung pada kualitas dan kelengkapan dokumen sumber yang digunakan dalam proses retrieval. Keterbatasan pada pengelolaan penyimpanan data internal berpotensi membatasi cakupan pengetahuan chatbot, khususnya pada pertanyaan yang memerlukan penggabungan informasi lintas konteks. Selain itu, aspek keamanan sistem belum sepenuhnya dioptimalkan karena belum diterapkannya mekanisme validasi permintaan dan pembatasan akses (*rate limiting*), yang berpotensi menyebabkan penyalahgunaan permintaan serta peningkatan konsumsi token pada layanan model generatif. Oleh karena itu, penelitian selanjutnya dapat difokuskan pada penguatan manajemen basis pengetahuan, optimalisasi mekanisme keamanan sistem, serta penerapan strategi kontrol akses untuk meningkatkan efisiensi dan keberlanjutan operasional chatbot berbasis RAG.

Deklarasi

Kontribusi Penulis. Semua penulis berkontrbusi secara bersama-sama dengan kontributor utama dalam artikel ini. Semua penulis membaca dan menyetujui versi akhir dari artikel yang diajukan.
Pernyataan Pendanaan. Tidak ada penulis yang menerima dana atau hibah dari lembaga atau badan pendanaan untuk penelitian ini.
Konflik Kepentingan. Penulis menyatakan tidak ada konflik kepentingan.
Informasi Tambahan. Tidak ada informasi tambahan dalam artikel ini.

Daftar Pustaka

[1] T.K. Yunianto, "Kata.ai: Penggunaan Chatbot Turunkan Biaya Operasional 70%," Accessed: January 26, 2026. [Online]. Available: https://www.marketeers.com/kata-ai-penggunaan-chatbot-turunkan-biaya-operasional-70/

[2] S. Wibawa, "Analisis Chatbot Otomatisasi Tugas Administratif dan Manajemen Dalam Lingkungan Digital Dengan Menggunakan Python," *Jurnal Inovasi dan Sains Teknik Elektro.,* vol. 4, no. 1, pp. 25-31, 2023.

[3] A. Haikal, I. Leliana, R. Septian, E. Kusnadi, D. Sad Tanti, "Pemanfaatan AI dan Chatbot dalam Praktik Public Relations: Adaptasi Teknologi dan Etika Komunikasi, " *Jurnal Public Relations-JPR.,* vol. 6, no. 1, pp. 62-67, 2025.

[4] A. Sahata Sitanggang, R. Fenny Syafariani, F. Wulan Sari, Wartika, N. Hasti, "Relation of Chatbot Usage Towards Customer Satisfaction Level in Indonesia, " *International Journal of Advances in Data and Information Systems.,* vol. 4, no. 1, pp. 86-96. 2023, Doi: 10.25008/ijadis.v4i1.1261.

[5] M. Roihan Racman, M. Rosidin, W. Yuli Sulistyo, "Implementasi Metode Retrieval Augmented Generation Pada Chatbot Untuk Otomatisasi Layanan Pelanggan Kontrakan, " *Informatics and security (insect).,* vol. 11, no. 2, pp. 229-237, 2025.

[6] I. Afandi, "Implementasi Chatbot sebagai Solusi Inovatif dalam Pelayanan Akademik, " *Journal of Cyber Health and Computer (JOCHAC).,* vol. 3, no. 1, pp. 25-29. 2025, Doi: 10.64163/jochac.v3i1.36.

[7] E. Lavrinovics, R. Biswas, J. Bjerva, K. Hose, "Knowledge Graphs, Large Language Models, and Hallucinations: An NLP, " *Web Semantics: Science, Services and Agents on the World.,* vol. 85, pp. 1-7. 2025, Doi: https://doi.org/10.1016/j.websem.2024.100844.

[8] T. Ramadhani, N. Qotrun Nada, N. Dwi S, "Penerapan Metode Retrieval-Augmented Generation (RAG) Pada Chatbot E-Commerce Berbasis Gemini Ai, " *Jurnal Ilmiah ILKOMINFO – Ilmu Komputer & Informatika.,* vol. 8, no. 2, pp. 301-313, 2025.

[9] M. Wysocka, O. Wysocki, M. Delmas, V. Mutel, A. Freitas, "Large Language Models, scientific knowledge and factuality: A framework to streamline human expert evaluation, " *Journal of Biomedical Informatics.,* vol. 158, pp. 1-19. 2024, Doi: https://doi.org/10.1016/j.jbi.2024.104724.

[10] Murtiyoso, I. Tahyudin, Berlilana, "A Systematic Review of Retrieval-Augmented Generation for Enhancing Domain-Specific Knowledge in Large Language Models, " *Sinkron : Jurnal dan Penelitian Teknik Informatika.,* vol. 9, no. 2, pp. 969-977. 2025, Doi: https://doi.org/10.33395/sinkron.v9i2.14824.

[11] I. Pujiono, I. Murtadho Agtyaputra, Y. Ruldeviyani, "Implementing Retrieval-Augmented Generation And Vector Databases For Chatbots In Public Services Agencies Context, " *JITK (Jurnal Ilmu Pengatahuan dan Komputer).,* vol. 10, no. 1, pp. 216-223. 2024, Doi: 10.33480/jitk.v10i1.5572.

[12] R.D. Andika Raya, F.L. Dwi Cahyanti, "Perancangan Sistem Informasi Chatbot Retrieval Augmented Generation Berbasis Website Pada PT. Revolusi Cita Edukasi, " *Indonesian Journal Computer Science.,* vol. 4, no. 1, pp. 15-21. 2025.

[13] G.D. Albert, A. Voutama, "Pengembangan Chatbot Berbasis Pdf Menggunakan Local Retrieval-Augmented Generation (RAG) dan Ollama, " *Jitet: Jurnal Informatika dan Teknik Elektro Terapan.*, vol. 13, no. 2, pp. 937-944. 2025, Doi: http://dx.doi.org/10.23960/jitet.v13i2.6361.

[14] M. Syarifudin, E. Yulianto, A. Nugroho L.I.F, "Modeling AI-Chatbot Service Quality and Purchase Intention: Mediating Mechanisms and the Moderating Role of Intrusiveness, " *Journal of Digital*

Marketing and Halal Industry., vol. 6, no. 2, pp. 211-240. 2024, Doi: https://doi.org/10.21580/jdmhi.2024.6.2.27893.

[15] M.D. Arya Muhajir, N. Prasiti, M. Koeshardianto, "Implementasi Chatbot Menggunakan Framework Langchain Berbasis LLM GPT (Studi Kasus : Panduan Akademik Universitas Trunojoyo), " *JATI (Jurnal Mahasiswa Teknik Informatika).,* vol. 9, no. 2, pp. 2151-2158. 2025.