Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

Optimasi Akurasi Jawaban Aplikasi Chatbot Layanan Pelanggan dengan Metode RAGRetrieval-Augmented Generation

Dhaman*, Sajarwo Anggai, Arya Adhyaksa Waskita

Magister Teknik Informatika, Universitas Pamulang, Tangerang Selatan
Jl. Raya Puspitek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten, Indonesia
Email: [1,*]mas.dhaman@gmail.com, [2]dosen02832@unpam.ac.id, [3]aawaskita@unpam.ac.id
Email Penulis Korespondensi: mas.dhaman@gmail.com
Submitted: **18/07/2025;** Accepted: **30/07/2025;** Published: **31/07/2025**

Abstrak−Penelitian ini membahas permasalahan rendahnya akurasi jawaban pada sistem chatbot berbasis Large Language Model (LLM) dalam menjawab pertanyaan dari dokumen layanan pelanggan. Untuk mengatasi hal tersebut, digunakan pendekatan metode Retrieval-Augmented Generation (RAG) guna meningkatkan kualitas jawaban dengan menambahkan konteks relevan dari dokumen eksternal. Tiga model LLM yang digunakan dalam penelitian ini adalah Llama3.1:8B, Llama3.2:1B, dan Llama3.2:3B dari Meta AI. Evaluasi dilakukan menggunakan metrik otomatis ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) serta penilaian manual oleh manusia (human evaluation) terhadap aspek akurasi, relevansi, dan halusinasi. Penelitian ini berkontribusi pada pengembangan sistem tanya-jawab berbasis LLM yang lebih handal dengan konteks dokumen eksternal dengan informasi seputar layanan pelanggan. Hasil penelitian menunjukkan bahwa penerapan metode RAG memberikan peningkatan yang signifikan terhadap seluruh model. Skor F1 ROUGE meningkat secara konsisten di semua model, dengan peningkatan tertinggi pada Llama3.1:8b (dari 0.12 menjadi 0.58 pada ROUGE-1). Evaluasi manual juga menunjukkan peningkatan akurasi (hingga +2.73 poin) dan penurunan halusinasi (hingga −2.63 poin). Peningkatan performa tidak hanya terjadi pada model besar, tetapi juga pada model kecil, membuktikan bahwa manfaat RAG tidak bergantung pada ukuran model. Kesimpulan dari penelitian ini adalah bahwa metode RAG sangat efektif dalam meningkatkan akurasi dan keandalan jawaban chatbot, khususnya dalam skenario tanya-jawab berbasis dokumen. Dengan memanfaatkan konteks dari dokumen eksternal, sistem dapat menghasilkan jawaban yang lebih faktual, relevan, dan minim halusinasi. RAG terbukti sebagai pendekatan yang efisien untuk meningkatkan kualitas jawaban LLM, termasuk dengan model LLM yang mempunya ukuran parameter lebih kecil.

Kata Kunci: Large Language Model; Llama; Retrieval-Augmented Generation; ROUGE; Evaluasi

Abstract−This research addresses the issue of low answer accuracy in chatbot systems based on Large Language Models (LLMs) when responding to questions derived from customer service documents. To overcome this problem, the Retrieval-Augmented Generation (RAG) method is applied to improve the quality of responses by adding relevant context from external documents. Three LLM models used in this study are LLaMA3.1 8B, LLaMA3.2 1B, and LLaMA3.2 3B from Meta AI. Evaluation is conducted using automatic ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) and manual human evaluation assessing accuracy, relevance, and hallucination. This research contributes to the development of more reliable question-answering systems based on LLMs enhanced with external contextual documents related to customer service information. The results show a significant improvement across all models after applying the RAG method. ROUGE F1-scores increased consistently, with Llama3.1:8b showing the highest gain (from 0.12 to 0.58 on ROUGE-1). Human evaluation also confirmed improvements in accuracy (up to +2.73 points) and reductions in hallucination (up to −2.63 points). These improvements were evident not only in larger models but also in smaller ones, indicating that the benefits of RAG are not dependent on model size. In conclusion, RAG is highly effective in enhancing the accuracy and reliability of chatbot responses, especially in document-based question-answering scenarios. By leveraging contextual information from external documents, the system produces more factual, relevant, and hallucination-free responses. RAG has proven to be an effective approach for enhancing the response quality of LLM, including those with smaller parameter sizes.

Keywords: Large Language Model; Llama; Retrieval-Augmented Generation; ROUGE; Evaluation

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI) telah membawa transformasi signifikan dalam pengolahan bahasa alami (Natural Language Processing/NLP). Salah satu terobosan besar adalah hadirnya Large Language Model (LLM) seperti GPT, Llama , dan Gemma yang mampu menghasilkan teks secara otomatis dengan konteks dan makna yang relevan [1]. Namun, meskipun LLM memiliki kemampuan bahasa yang luar biasa, terdapat tantangan besar dalam hal akurasi jawaban, terutama saat LLM digunakan dalam skenario open-domain question answering atau sistem tanya jawab berbasis dokumen.

Masalah utama yang dihadapi dalam penggunaan LLM adalah kecenderungan model menghasilkan jawaban yang halusinatif atau tidak sesuai dengan sumber data yang valid [2]. Hal ini disebabkan oleh sifat dasar LLM yang mengandalkan parameter yang telah dilatih pada data besar tanpa akses langsung ke informasi eksternal atau kontekstual secara real-time. Ketika diminta menjawab pertanyaan berbasis dokumen atau domain tertentu, LLM sering kali tidak dapat memberikan jawaban yang tepat atau aktual [3].

Untuk mengatasi permasalahan tersebut, pendekatan Retrieval-Augmented Generation (RAG) diperkenalkan sebagai solusi yang menjanjikan [4]. RAG menggabungkan kemampuan pencarian informasi (retrieval) dengan generasi jawaban oleh LLM. Dalam pendekatan ini, sistem melakukan pengambilan informasi dari sumber eksternal yang relevan (seperti dokumen, webpage atau basis data), kemudian hasil pengambilan

Copyright © 2025 The Author, Page 2120
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

tersebut dijadikan konteks tambahan bagi LLM untuk menghasilkan jawaban. Dengan demikian, proses ini diharapkan dapat meningkatkan akurasi dan relevansi jawaban yang diberikan oleh model [5].

Penelitian mengenai RAG telah berkembang dalam beberapa tahun terakhir. Lewis et al. (2020) memperkenalkan RAG dalam konteks model generatif dan menunjukkan peningkatan performa signifikan pada beberapa dataset seperti Natural Questions dan WebQuestions [6]. Sementara itu, Karpukhin et al. (2020) mengusulkan Dense Passage Retrieval (DPR) sebagai komponen retriever yang efektif untuk tugas tanya jawab berbasis dokumen [7]. Studi terbaru oleh Wang et al. (2023) menyoroti pentingnya pemilihan dokumen dan strategi pemrosesan multi-hop dalam konteks RAG untuk meningkatkan akurasi jawaban pada domain ilmiah dan teknis [8].

Meskipun berbagai pendekatan RAG telah terbukti efektif di banyak studi [9][10], namun masih terdapat celah penelitian (research gap) dalam penerapan metode RAG, khususnya pada penggunaan dokumen eksternal berbahasa Indonesia dan pemanfaatan model LLM ringan seperti Llama3.2 dengan ukuran parameter 1B dan 3B. Celah ini membuka peluang untuk dijadikan objek penelitian lebih lanjut, mengingat relevansinya dalam pengembangan aplikasi LLM lokal yang menghadapi keterbatasan perangkat keras serta kebutuhan untuk menjaga privasi data. Oleh karena itu, implementasi chatbot berbasis LLM dalam konteks lokal menjadi tantangan sekaligus peluang untuk mengembangkan sistem tanya-jawab yang efisien, akurat, dan aman. Penelitian terkait penerapan RAG di konteks lokal dan bahasa non-Inggris, termasuk Bahasa Indonesia, masih sangat terbatas. Padahal, pengujian terhadap dataset spesifik seperti dokumen ilmiah nasional, jurnal kesehatan lokal, artikel hukum Indonesia atau dokumen spesifik lainnya sangat dibutuhkan untuk mendukung pengembangan solusi berbasis AI yang lebih inklusif dan relevan secara lokal.

Sebagian besar penelitian sebelumnya dilakukan pada dataset berbahasa Inggris seperti Natural Questions, TriviaQA, HotpotQA, atau BioASQ, yang mewakili konteks dan struktur informasi dari negara-negara Barat. Sebagai contoh, penelitian oleh Lewis et al. (2020) dan Karpukhin et al. (2020) menunjukkan keberhasilan RAG dalam menjawab pertanyaan dari dokumen Wikipedia dan artikel berita berbahasa Inggris [6][7] . Namun, struktur penulisan, gaya bahasa, serta keragaman terminologi dalam dokumen-dokumen lokal di Indonesia, seperti jurnal ilmiah berbahasa Indonesia, dokumen kebijakan pemerintah, atau laporan penelitian kesehatan, sangat berbeda dengan dokumen-dokumen global tersebut [11]. Oleh karena itu, pendekatan dan model yang terbukti efektif di luar negeri belum tentu menghasilkan performa yang optimal jika langsung diterapkan di konteks Indonesia tanpa penyesuaian.

Penelitian ini bertujuan untuk mengoptimalkan akurasi jawaban LLM dengan mengimplementasikan metode RAG, khususnya dalam skenario sistem tanya jawab berbasis dokumen berbahasa Indonesia. Sistem yang dikembangkan akan menggunakan pipeline RAG yang terdiri dari proses splitting dokumen, embedding chunk teks, serta retrieval menggunakan similarity search. Model LLM ringan Llama 3.1 dan Llama 3.2 dari Meta digunakan agar dapat diuji secara lokal. Evaluasi sistem dilakukan menggunakan metrik otomatis ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) serta dilengkapi dengan evaluasi manual menggunakan human evaluation untuk menilai relevansi, akurasi, dan halusinasi jawaban.

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem LLM yang akurat, efisien, dan dapat diimplementasikan secara lokal tanpa tergantung pada layanan cloud atau perangkat komputer dengan resource yang besar. Selain itu, studi ini juga diharapkan dapat menjadi dasar untuk penelitian lanjutan di bidang LLM dengan pendekatan metode RAG dan aplikasinya dalam bahasa-bahasa selain Inggris, khususnya dalam konteks dokumen ilmiah, question answering bisnis perusahaan dan pendidikan di Indonesia. Penelitian ini juga membuka peluang bagi pemanfaatan model-model LLM dengan jumlah parameter yang lebih kecil (lightweight LLM) agar tetap mampu memberikan jawaban yang relevan dan berkualitas tinggi ketika dikombinasikan dengan teknik retrieval yang tepat. Dengan pendekatan RAG, keterbatasan kapasitas model dapat dikompensasi melalui penyediaan konteks eksternal yang relevan, sehingga model kecil tetap dapat berfungsi secara optimal dalam skenario dunia nyata yang memerlukan akurasi informasi.

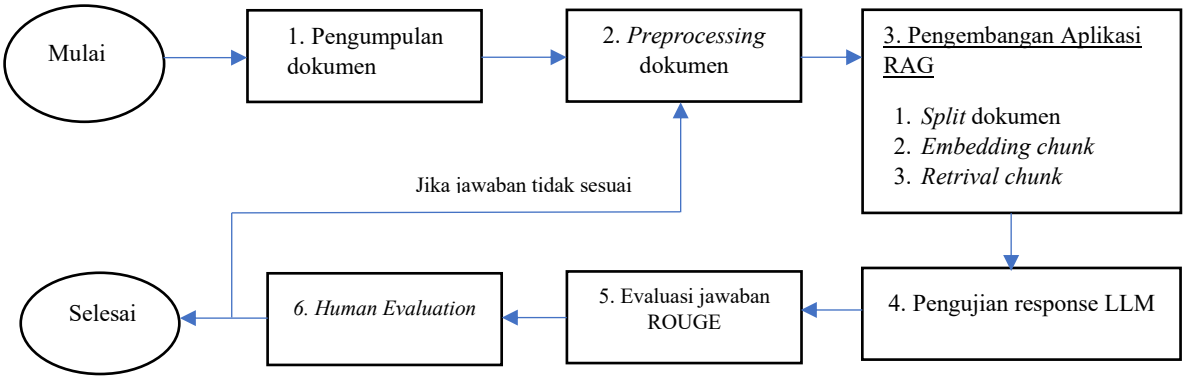
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan sistematis untuk mengembangkan dan mengevaluasi sistem chatbot berbasis LLM yang diperkuat dengan metode RAG. **Gambar 1** berikut ini adalah tahapan dalam penelitian yang menjelaskan secara umum alur proses penelitian mulai dari pengumpulan dokumen hingga evaluasi akhir oleh pengguna. Setiap tahapan dirancang untuk memastikan bahwa sistem yang dibangun mampu menghasilkan jawaban yang relevan, akurat, dan minim halusinasi terhadap pertanyaan dari dokumen layanan pelanggan.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048



Gambar 1. Tahapan Penelitian

Penelitian ini dilakukan dengan serangkaian tahapan sistematis untuk menguji efektivitas metode RAG dalam meningkatkan akurasi jawaban model LLM. Tahapan dimulai dengan proses pengumpulan dokumen(1), yaitu mengumpulkan sejumlah data teks antara data dokumen profil perusahaan, katalog produk dan panduan layanan pelanggan. Dokumen ini dikumpulkan dari bagian layanan pelanggan perusahaan dan dari website resmi perusahaan. Dokumen tersebut menjadi dasar dalam proses pencarian konteks ketika LLM diminta untuk menjawab pertanyaan.

Langkah selanjutnya adalah preprocessing dokumen(2), di mana seluruh dokumen yang telah dikumpulkan akan dibersihkan dari jenis data selain teks, dimana pada penelitian hanya data teks yang akan diproses. Selanjutnya dokumen ini akan diformat sesuai dengan panduan dari Docling selaku library untuk split dokumen [12]. Format ini menggunakan format h2 (header 2) untuk judul atau section dan format paragraph untuk seluruh isi yang ada dalam section tersebut.

Setelah preprocessing selesai, tahapan dilanjutkan adalah pengembangan aplikasi(3) yang akan menjalankan RAG Pipeline [13], yang terdiri dari tiga proses utama yaitu pemotongan dokumen menjadi chunk menggunakan Docling, konversi setiap chunk ke dalam bentuk vektor menggunakan OllamaEmbeddings dengan model embedding adalah nomic-embed-text untuk selanjutnya disimpan dalam pgvector di PostgreSQL [14], tahap selanjutnya melakukan proses pencarian (retrieval) menggunakan teknik semantik similarity untuk menemukan potongan teks yang paling relevan terhadap pertanyaan yang diajukan [15].

Hasil retrieval dari pipeline RAG akan digunakan dalam pengujian response LLM(4), yaitu proses di mana potongan dokumen relevan diberikan sebagai konteks tambahan ke dalam prompt LLM untuk menghasilkan jawaban. Dalam penelitian ini, model LLM yang digunakan adalah model LLM dari Meta yaitu Llama3.1 dan Llama3.2 [16] yang dioperasikan secara lokal melalui provider Ollama [9]. Proses terakhir adalah melakukan evaluasi terhadap hasil yang dikeluarkan oleh LLM dengan menggunakan metrik otomatis ROUGE(5) dan evaluasi manual oleh pengguna atau human evaluation(6).

2.2 Evaluasi

Evaluasi dalam penelitian ini bertujuan untuk mengukur akurasi dan relevansi jawaban yang dihasilkan oleh model LLM dalam menjawab pertanyaan berbasis dokumen. Evaluasi dilakukan dengan dua pendekatan utama, yaitu evaluasi kuantitatif menggunakan metrik ROUGE [17] dan evaluasi kualitatif melalui penilaian manual oleh manusia (human evaluation).

2.2.1 Evaluasi Kuantitatif dengan ROUGE

Evaluasi otomatis dilakukan menggunakan metrik ROUGE (Recall-Oriented Understudy for Gisting Evaluation), yang merupakan metrik evaluasi populer dalam bidang NLP untuk mengukur kesamaan antara teks prediksi dan teks referensi [17]. Dalam penelitian ini digunakan tiga jenis metrik ROUGE, yaitu ROUGE-1 untuk menghitung kesamaan uni-gram antara jawaban prediksi dengan referensi, ROUGE-2 untuk menghitung kesamaan bigram dan ROUGE-L untuk mengukur Longest Common Subsequence (LCS) untuk menghitung kesamaan dalam urutan kalimat.

Metrik ROUGE ini digunakan untuk menghitung nilai presisi atau ketepatan proporsi kata yang benar terhadap seluruh kata dalam jawaban prediksi, yang kedua untuk menghitung recall atau kelengkapan yaitu kata yang benar terhadap seluruh kata pada jawaban referensi dan yang ketiga untuk menghitung F1-Score yaitu rata-rata harmonis antara presisi dan recall.

a. Precision

$$\text{Precision}_{\text{(Rouge-n)}} = \frac{\text{Jumlah n−gram kata sama}}{\text{Keseluruhan kata di jawaban LLM}} \quad (1)$$

$$\text{Precision}_{\text{(Rouge-L)}} = \frac{\text{LCS (Longest Common Subsquent)}}{\text{Keseluruhan kata di jawaban LLM}} \quad (2)$$

Copyright © 2025 The Author, Page 2122
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

b. Recall

$$\text{Recall}_{\text{(Rouge-n)}} = \frac{\text{Jumlah n−gram kata sama}}{\text{Keseluruhan kata di referensi}} \quad (3)$$

$$\text{Recall}_{\text{(Rouge-L)}} = \frac{\text{LCS (Longest Common Subsquent)}}{\text{Keseluruhan kata di referensi}} \quad (4)$$

c. F1 Score

$$\text{F1 Score}_{\text{(Rouge-n)}} = 2 \text{ x } \frac{\text{Precision}_{\text{(Rouge−n)}} \text{ x Recall}_{\text{(Rouge−n)}}}{\text{Precision}_{\text{(Rouge−n)}} + \text{Recall}_{\text{(Rouge−n)}}} \quad (5)$$

$$\text{F1 Score}_{\text{(Rouge-L)}} = 2 \text{ x } \frac{\text{Precision}_{\text{(Rouge−L)}} \text{ x Recall}_{\text{(Rouge−L)}}}{\text{Precision}_{\text{(Rouge−L)}} + \text{Recall}_{\text{(Rouge−L)}}} \quad (6)$$

2.2.2 Evaluasi Kualitatif dengan Human Evaluation

Sebagai pelengkap dari evaluasi metrik otomatis, dilakukan pula evaluasi kualitatif melalui penilaian manusia (human evaluation). Tujuan dari evaluasi ini adalah untuk mengamati aspek-aspek yang tidak dapat ditangkap oleh metrik numerik, seperti relevansi, akurasi dan halusinasi [18] [19].

Proses evaluasi dilakukan dengan memberikan sejumlah pertanyaan dan jawaban dari model LLM kepada evaluator manusia, baik pada kondisi dengan penerapan RAG maupun tanpa RAG. Evaluator kemudian menilai kualitas jawaban berdasarkan tiga aspek utama: (1) Relevansi, yakni sejauh mana jawaban sesuai dengan isi dokumen; (2) Akurasi, yakni tingkat kebenaran informasi dalam jawaban; (3) Halusinasi, yakni sejauh mana jawaban mengandung informasi yang tidak ada atau menyimpang dari dokumen sumber.

Penilaian dilakukan menggunakan skala Likert 1 sampai 5 untuk masing-masing aspek [20], untuk aspek akurasi dan relevansi, skor 1 menunjukkan kualitas sangat buruk dan skor 5 menunjukkan kualitas sangat baik, sedangkan untuk aspek halusinasi skor 1 menunjukkan kualitas yang sangat baik sedangkan skor 5 menunjukkan kualitas yang buruk. Setiap pertanyaan dinilai oleh lebih dari satu evaluator untuk memastikan reliabilitas hasil penilaian. Data dari hasil evaluasi kemudian diolah untuk mendapatkan nilai rata-rata per kriteria serta dibandingkan antara model.

3. HASIL DAN PEMBAHASAN

3.1 RAG Pipeline

RAG pipeline terdiri dari tiga komponen utama, yaitu proses pemotongan dokumen (split), pembentukan representasi vektor (embedding), dan pencarian informasi yang relevan (retrieval). Ketiga tahapan ini membentuk alur utama yang mendukung sistem LLM dalam menghasilkan jawaban berbasis dokumen.

3.1.1 Documents Split

Tahapan pertama adalah proses split dokumen, yaitu memecah dokumen panjang menjadi potongan-potongan kecil (chunk) yang lebih mudah untuk diproses dan dibandingkan secara semantik. Dalam penelitian ini digunakan fungsi RecursiveCharacterTextSplitter dari LangChain yang memotong teks berdasarkan struktur kalimat atau paragraf [21] dengan maksimal ukuran chunk adalah 1000 kata dan overlap 200 kata. Proses ini memastikan bahwa setiap potongan dokumen tetap memiliki makna utuh dan kontekstual saat dilakukan embedding dan pencarian. Pada tahap split dokumen, seperti yang ditunjukkan pada Tabel 1, dokumen Profil Perusahaan(1) menghasilkan sebanyak 23 potongan chunk, dokumen Katalog(2) Produk menghasilkan 79 potongan chunk, sedangkan dokumen Layanan Pelanggan(3) menghasilkan 69 potongan chunk.

Tabel 1. Jumlah chunks hasil dari splitting dokumen

No	Nama Dokumen	Banyaknya chunk
1	Profil perusahaan	23
2	Katalog produk	79
3	Layanan pelanggan	69

3.1.2 Embedding

Setelah dokumen dipotong, setiap chunk diubah menjadi vektor numerik berdimensi tinggi melalui proses embedding. Embedding memungkinkan teks direpresentasikan dalam bentuk vektor sehingga dapat dibandingkan secara matematis menggunakan jarak semantik (menggunakan cosine similarity) [15]. Model embedding yang digunakan dalam penelitian ini adalah nomic-embed-text yang dijalankan secara lokal melalui layanan Ollama. Setiap chunk yang telah diproses embedding akan disimpan dalam basis data vektor menggunakan pgvector pada PostgreSQL. Penyimpanan ini memungkinkan sistem melakukan pencarian vektor secara efisien menggunakan teknik similarity.

3.1.3 Retrieval

Copyright © 2025 The Author, Page 2123
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

Tahap selanjutnya adalah retrieval, yaitu proses pencarian potongan dokumen paling relevan terhadap pertanyaan pengguna. Sistem menggunakan similarity search untuk membandingkan embedding dari pertanyaan dengan seluruh vektor chunk yang tersimpan [22][23]. Chunk dengan nilai kemiripan tertinggi akan dipilih sebagai konteks yang diberikan ke LLM.

Dalam implementasi ini digunakan metode pencarian berbasis cosine similarity melalui pgvector, dan sistem hanya mengambil 4 potongan teratas untuk dijadikan konteks jawaban. Dengan teknik ini, RAG memungkinkan model LLM diharapkan akan menghasilkan jawaban yang lebih relevan terhadap dokumen, bahkan jika dokumen sangat panjang atau tidak terstruktur.

3.2 Pengujian

Percobaan dalam penelitian ini dilakukan menggunakan dua metode evaluasi utama. Metode pertama adalah evaluasi kuantitatif menggunakan metrik ROUGE, yang membandingkan jawaban yang dihasilkan oleh model LLM dengan jawaban referensi yang diperoleh dari dokumen sumber. Metrik ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) digunakan untuk mengukur tingkat kesamaan n-gram dan kesesuaian struktur antara jawaban prediksi dan jawaban ideal.

Metode kedua adalah evaluasi kualitatif melalui penilaian manusia (human evaluation). Evaluasi ini dilakukan oleh beberapa evaluator untuk menilai akurasi konten dan tingkat halusinasi dari jawaban yang dihasilkan. Penilaian dilakukan dengan mempertimbangkan relevansi jawaban terhadap dokumen sumber, kebenaran informasi, serta kesesuaian jawaban dengan pertanyaan yang diajukan. Metode ini bertujuan untuk melengkapi evaluasi metrik otomatis dengan pendekatan subjektif yang dapat menangkap kualitas jawaban dari perspektif pemahaman manusia.

Pengujian pertama adalah perbandingan metrik ROUGE menggunakan metode RAG pada dokumen profil perusahaan dengan jumlah percobaan sebanyak 32 pertanyaan yang diambil rata-rata dari masing-masing model LLM, Tabel 2 berikut ini menampilkan skor hasil evaluasi pada dokumen profil perusahaan.

Tabel 2. Metrik ROUGE menggunakan metode RAG pada dokumen Profil Perusahaan

MODEL	ROUGE 1			ROUGE 2			ROUGE L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Llama3.1:8b	0,045138889	0,042361111	0,041666667	00.58	00.51	00.52	0,04375	00.59	00.58
Llama3.2:1b	00.35	00.39	00.33	00.21	00.25	00.22	00.33	00.36	00.31
Llama3.2:3b	00.43	00.47	00.40	00.33	00.32	00.30	00.41	00.44	00.39

Tabel 2 menunjukkan bahwa model Llama3.1:8B menghasilkan performa terbaik pada semua metrik ROUGE, baik ROUGE-1, ROUGE-2, maupun ROUGE-L, dengan skor F1 yang paling tinggi dibandingkan model lainnya. Sementara itu, model Llama3.2:1B menunjukkan performa terendah, sedangkan Llama3.2:3B berada di posisi tengah dengan hasil yang lebih baik dari 1B namun belum melampaui 8B.

Pengujian kedua adalah perbandingan metrik ROUGE menggunakan metode RAG pada dokumen katalog produk dengan jumlah percobaan sebanyak 74 pertanyaan yang diambil rata-rata dari masing-masing model LLM. Tabel 3 berikut ini menampilkan skor hasil evaluasi pada dokumen katalog produk.

Tabel 3. Metrik ROUGE menggunakan metode RAG pada dokumen katalog produk

model LLM	ROUGE 1			ROUGE 2			ROUGE L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Llama3.1:8b	0,04375	0,047916667	0,041666667	00.52	00.57	00.50	0,041666667	0,045833333	00.57
Llama3.2:1b	00.39	00.53	00.40	00.25	00.35	00.26	00.34	00.47	00.34
Llama3.2:3b	00.59	0,051388889	0,042361111	00.47	0,041666667	00.49	00.56	0,049305556	00.58

Berdasarkan hasil pada Tabel 3, terlihat bahwa model Llama3.2:3B dan Llama3.1:8b menunjukkan performa terbaik dalam menghasilkan jawaban berbasis dokumen katalog produk, dengan skor yang tidak terpaut jauh antara keduanya. Sementara itu, model Llama3.2:1B kembali menunjukkan performa paling rendah di semua metrik, mengindikasikan keterbatasannya dalam memahami konteks meskipun dibantu metode RAG. Menariknya, pada jenis dokumen ini, model 3B berhasil melampaui model 8B pada hampir seluruh metrik, yang menunjukkan bahwa dengan struktur dokumen tertentu, model berukuran sedang dapat lebih efektif dibandingkan model besar.

Pengujian ketiga adalah perbandingan metrik ROUGE menggunakan metode RAG pada dokumen layanan pelanggan dengan jumlah percobaan sebanyak 66 pertanyaan yang diambil rata-rata dari masing-masing model LLM, hasil evaluasi terlihat pada Tabel 4 berikut ini.

Tabel 4. Metrik ROUGE menggunakan metode RAG pada dokumen layanan pelanggan

model LLM	ROUGE 1			ROUGE 2			ROUGE L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Llama3.1:8b	**0,04236111**	00.56	00.54	**00.53**	00.48	00.46	**0,04166667**	00.54	00.52
Llama3.2:1b	00.29	00.55	00.35	00.21	00.41	00.25	00.25	00.49	00.31
Llama3.2:3b	00.41	0,042361111	00.46	00.33	00.51	00.38	00.40	00.59	00.44

Copyright © 2025 The Author, Page 2124
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

Berdasarkan hasil pada Tabel 4, model Llama3.1:8B menunjukkan performa terbaik dalam menghasilkan jawaban terhadap dokumen layanan pelanggan dibandingkan dengan 2 model lainnya, dengan skor F1 tertinggi pada semua metrik ROUGE, yaitu ROUGE-1 sebesar 0.54, ROUGE-2 sebesar 0.46, dan ROUGE-L sebesar 0.52. Model Llama3.2:1B mencatat performa terendah secara keseluruhan, dengan skor F1 ROUGE-1 hanya sebesar 0.35, ROUGE-2 sebesar 0.25, dan ROUGE-L sebesar 0.31.

Pengujian keempat adalah perbandingan metrik ROUGE tanpa menggunakan metode RAG, mengambil jawaban langsung dari LLM dengan jumlah percobaan sebanyak 172 pertanyaan yang diambil rata-rata dari masing-masing model LLM, hasil evaluasi dapat dilihat pada Tabel 5 berikut ini.

Tabel 5. Metrik ROUGE tanpa menggunakan metode RAG

model LLM	ROUGE 1			ROUGE 2			ROUGE L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Llama3.1:8b	00.22	00.10	00.12	00.06	00.03	00.04	00.20	00.09	00.11
Llama3.2:1b	00.16	00.16	00.14	00.03	00.04	00.03	00.13	00.14	00.12
Llama3.2:3b	00.19	00.15	00.15	00.05	00.04	00.03	00.16	00.13	00.12

Pada Tabel 5, terlihat bahwa performa ketiga model LLM secara umum sangat rendah ketika tidak menggunakan metode RAG. Skor F1 pada seluruh metrik ROUGE hanya berkisar antara 0.03 hingga 0.15, menunjukkan bahwa model kesulitan menghasilkan jawaban yang relevan dan akurat tanpa adanya bantuan konteks dari dokumen eksternal. Tidak ada model yang secara signifikan unggul, meskipun Llama3.2:1B dan Llama3.2:3B sedikit lebih stabil pada beberapa metrik dibandingkan Llama3.1:8B.

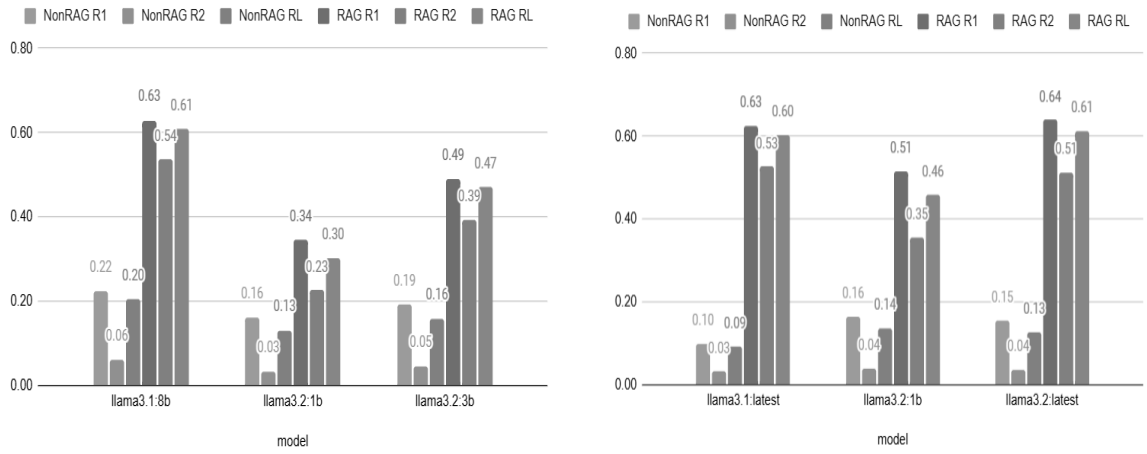
Gabungan semua percobaan dengan RAG yaitu dengan dokumen profil Perusahaan, katalog produk dan layanan pelanggan, hasil dari semua percobaan nilainya dirata-rata, Tabel 6 berikut ini menunjukkan skor ROUGE gabungan dari semua dokumen.

Tabel 6. Metrik ROUGE menggunakan metode RAG semua dokumen

model LLM	ROUGE 1			ROUGE 2			ROUGE L		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Llama3.1:8b	0,04375	0,043055556	00.58	00.54	00.52	00.50	0,042361111	0,041666667	00.56
Llama3.2:1b	00.34	00.49	00.36	00.22	00.34	00.24	00.31	00.44	00.32
Llama3.2:3b	00.48	0,041666667	00.49	00.38	00.48	00.39	00.46	00.58	00.47

Berdasarkan Tabel 6 diatas yang merupakan gabungan hasil evaluasi ROUGE dari ketiga jenis dokumen (profil perusahaan, katalog produk, dan layanan pelanggan), model Llama3.1:8B secara konsisten mencatat skor tertinggi pada seluruh metrik, dengan F1-Score ROUGE-1 sebesar 0.58, ROUGE-2 sebesar 0.50, dan ROUGE-L sebesar 0.56. Model Llama3.2:3B menunjukkan performa menengah dengan F1-Score masing-masing 0.49, 0.39, dan 0.47, yang mengindikasikan bahwa meskipun tidak seakurat model 8B, ia tetap memberikan hasil yang kompetitif. Sementara itu, model Llama3.2:1B berada di posisi terbawah dengan skor F1 yang cukup rendah di seluruh metrik. Kesimpulan dari hasil ini menunjukkan bahwa penggunaan metode RAG secara signifikan meningkatkan kualitas jawaban untuk semua model, namun besarnya parameter model tetap menjadi faktor penting dalam menghasilkan output yang lebih relevan dan akurat. Model berukuran sedang seperti Llama3.2:3B dapat menjadi alternatif yang layak jika terdapat keterbatasan sumber daya, namun model besar seperti Llama3.1:8B masih menjadi pilihan utama untuk performa maksimal.

Untuk memperoleh gambaran yang lebih jelas mengenai dampak penerapan metode RAG terhadap kualitas jawaban yang dihasilkan oleh model LLM, dilakukan analisis visual melalui grafik perbandingan metrik ROUGE. Gambar 2 berikut ini menyajikan perbandingan nilai Precision, Recall dari tiga varian model Llama baik dalam kondisi tanpa RAG (non-RAG) maupun dengan RAG.



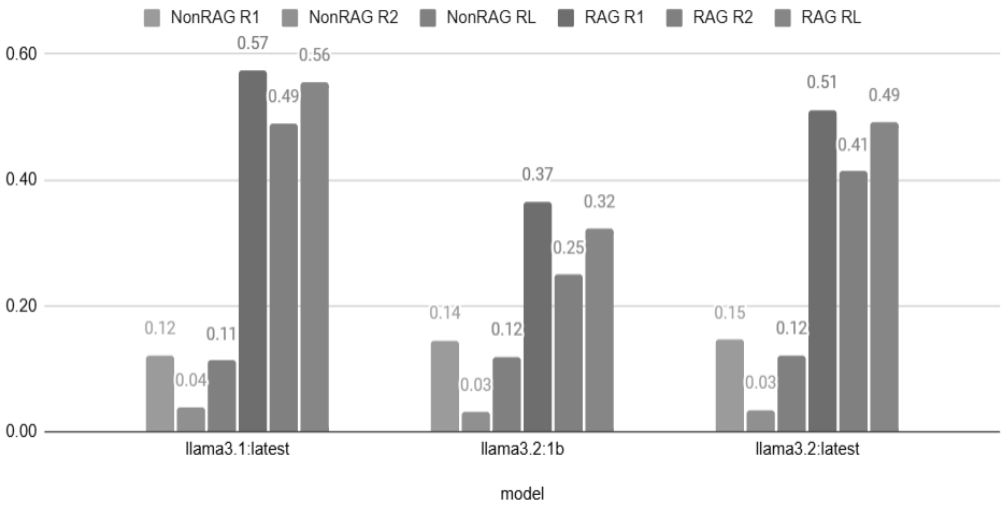
Gambar 2. Grafik Precision dan Grafik Recall

Copyright © 2025 The Author, Page 2125
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

Setiap grafik menunjukkan peningkatan yang signifikan pada semua metrik evaluasi ketika RAG digunakan, terutama pada model Llama3.1:8b dan Llama3.2:3b. Peningkatan ini mengindikasikan bahwa penambahan konteks melalui retrieval dokumen relevan dapat membantu LLM menghasilkan jawaban yang lebih akurat dan relevan terhadap pertanyaan yang diajukan.

Selanjutnya grafik skor F1 juga ditampilkan untuk mendapatkan nilai harmonisasi antara skor precision dan recall. Gambar 2 berikut ini menunjukkan hasil rata-rata dari seluruh percobaan yang dilakukan, baik tanpa menggunakan RAG maupun dengan menggunakan RAG.



Gambar 3. Grafik F1-Score

Selain evaluasi otomatis menggunakan metrik ROUGE, penelitian ini juga dilengkapi dengan evaluasi kualitatif yang dilakukan oleh evaluator manusia (human evaluation). Tujuan dari evaluasi ini adalah untuk menangkap dimensi kualitas jawaban yang tidak dapat diukur secara numerik oleh metrik otomatis [19]. Evaluasi dilakukan terhadap jawaban yang dihasilkan oleh model LLM, baik pada kondisi tanpa RAG maupun dengan RAG.

Terdapat tiga aspek utama yang dinilai oleh evaluator, yaitu: **akurasi**, **relevansi** dan **halusinasi**. Masing-masing aspek dinilai dalam rentang skala Likert 1 sampai 5. Pada aspek akurasi dan relevansi, skor yang lebih tinggi menunjukkan kualitas jawaban yang semakin baik. Namun, khusus pada aspek halusinasi, skor yang lebih tinggi mengindikasikan tingkat kesalahan informasi yang semakin tinggi dan oleh karena itu bersifat negatif terhadap kualitas jawaban. Untuk menjaga konsistensi dalam perhitungan rata-rata kualitas, skor halusinasi perlu dibalik saat analisis, sehingga nilai rendah menunjukkan kualitas yang lebih baik.

Tabel 7 menunjukkan hasil evaluasi manual (human evaluation) yang dilakukan sebanyak 172 kali untuk masing-masing model LLM. Nilai-nilai yang ditampilkan merupakan rata-rata dari penilaian terhadap tiga aspek, yaitu akurasi, relevansi, dan halusinasi, pada ketiga jenis dokumen profil perusahaan, katalog produk, dan layanan pelanggan.

Tabel 7. Perbandingan human evaluation menggunakan metode RAG pada semua dokumen

model LLM	Profil Perusahaan			Katalog produk			QA Layanan Pelanggan		
	Akurasi	Relevansi	Halusinasi	Akurasi	Relevansi	Halusinasi	Akurasi	Relevansi	Halusinasi
Llama3.1:8b	0,19	0,17	0,10	0,20	0,22	0,07	0,19	0,21	0,05
Llama3.2:1b	0,14	0,15	0,12	0,17	0,19	0,10	0,18	0,19	0,10
Llama3.2:3b	0,17	0,18	0,10	0,21	0,22	0,06	0,19	0,21	0,06

Berdasarkan Tabel 7, model Llama3.1:8B secara konsisten memperoleh skor tertinggi dalam aspek akurasi dan relevansi, serta nilai halusinasi terendah pada seluruh jenis dokumen, menandakan kualitas jawaban yang paling faktual dan sesuai konteks. Model Llama3.2:3B menunjukkan performa yang mendekati model 8B, bahkan mencatat nilai akurasi tertinggi untuk dokumen katalog produk (4.62) dan nilai relevansi hampir setara, meskipun tingkat halusinasinya sedikit lebih tinggi dibanding Llama3.1:8B. Sementara itu, model Llama3.2:1B memiliki skor paling rendah pada semua aspek dan menunjukkan tingkat halusinasi tertinggi, terutama pada dokumen layanan pelanggan (1.82).

Melalui penilaian yang dilakukan oleh evaluator manusia, dapat diperoleh perbandingan yang lebih mendalam antara performa model LLM pada skenario tanpa RAG dan dengan penerapan RAG. Tabel 8 menyajikan hasil perbandingan penilaian manual terhadap ketiga model LLM dalam dua kondisi, yaitu tanpa menggunakan metode RAG dan dengan menggunakan metode RAG. Evaluasi mencakup tiga aspek utama: akurasi, relevansi, dan halusinasi, serta disertakan pula nilai perubahan (delta) dari masing-masing aspek untuk menunjukkan dampak langsung penerapan metode RAG.

Copyright © 2025 The Author, Page 2126
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

Tabel 8. Perbandingan nilai human evaluation tanpa RAG dan menggunakan RAG

model LLM	Tanpa RAG			Dengan RAG			Perubahan Nilai		
	Akurasi	Relevansi	Halusinasi	Akurasi	Relevansi	Halusinasi	Akurasi	Relevansi	Halusinasi
Llama3.1:8b	0,08	0,11	0,17	0,19	0,21	0,07	+2.73	+2.67	−2.63
Llama3.2:1b	0,10	0,09	0,19	0,18	0,18	0,11	+2.01	+2.08	−2.00
Llama3.2:3b	0,10	0,10	0,19	0,19	0,21	0,07	+2.45	+2.40	−2.54

Berdasarkan Tabel 8 diatas, seluruh model mengalami peningkatan signifikan dalam aspek akurasi dan relevansi setelah menggunakan metode RAG. Model Llama3.1:8B mencatat peningkatan tertinggi dengan selisih akurasi sebesar +2.73 poin dan relevansi sebesar +2.67 poin. Diikuti oleh Llama3.2:3B (+2.45 akurasi, +2.40 relevansi), serta Llama3.2:1B dengan peningkatan yang sedikit lebih rendah (+2.01 akurasi, +2.08 relevansi). Selain itu, terjadi penurunan tingkat halusinasi secara signifikan pada semua model, yang menunjukkan bahwa metode RAG berhasil mengurangi jawaban yang tidak faktual, dengan penurunan terbesar terjadi pada Llama3.1:8B (-2.63 poin). Temuan ini menegaskan bahwa integrasi metode RAG tidak hanya meningkatkan kualitas jawaban, tetapi juga secara efektif mengurangi risiko halusinasi pada sistem LLM, bahkan untuk model dengan parameter yang lebih kecil.

3.3 Pembahasan

Evaluasi performa model LLM terhadap kualitas jawaban dilakukan dengan menggunakan metrik ROUGE yang mencakup ROUGE-1, ROUGE-2, dan ROUGE-L. Masing-masing metrik diukur berdasarkan tiga aspek yaitu precision, recall, dan F1-score. Hasil evaluasi ini ditampilkan pada Tabel 5 (tanpa RAG) dan Tabel 6 (dengan RAG).

Secara umum, terdapat peningkatan yang sangat signifikan pada seluruh metrik ROUGE ketika metode RAG diterapkan. Hal ini menunjukkan bahwa penambahan konteks dari dokumen relevan berhasil meningkatkan kualitas dan kedekatan jawaban model terhadap referensi.

a. Model Llama3.1:8b
   Tanpa RAG skor F1 ROUGE-1 hanya 0.12, ROUGE-2 sebesar 0.04, dan ROUGE-L sebesar 0.11. Dengan RAG: Skor meningkat drastis menjadi 0.58 (ROUGE-1), 0.50 (ROUGE-2), dan 0.56 (ROUGE-L). Ini menunjukkan bahwa model kecil sekalipun dapat menunjukkan peningkatan besar ketika dibantu retrieval kontekstual.
b. Model Llama3.2:1b
   Tanpa RAG skor F1 ROUGE-1 hanya 0.14, ROUGE-2 sebesar 0.03 dan ROUGE-L sebesar 0.12. Dengan RAG skor F1 ROUGE-1 naik ke 0.36, ROUGE-2 sebesar 0.24, sedangkan ROUGE-L naik ke angka 0.32. Walaupun peningkatannya tidak sebesar model lain, namun tetap menunjukkan pengaruh positif dari RAG.
c. Model Llama3.2:3b
   Tanpa RAG skor F1 ROUGE-1 hanya 0.15, ROUGE-2 sebesar 0.03 dan ROUGE-L di angka 0.12. Sedangkan jika dilakukan ujicoba dengan RAG, Peningkatan terlihat jelas, yaitu diangka ROUGE-1 sebesar 0.49, ROUGE-2 sebesar 0.39 dan ROUGE-L sebesar 0.47.

Peningkatan skor F1 secara konsisten pada seluruh model dan metrik mengindikasikan bahwa metode RAG secara nyata meningkatkan kualitas jawaban model LLM baik dari segi kelengkapan (recall) maupun kesesuaian (precision). Hasil ini menguatkan argumen bahwa akurasi jawaban model tidak hanya bergantung pada ukuran model, tetapi juga pada konteks eksternal yang disediakan melalui retrieval dokumen. Hal ini juga dikuatkan dengan evaluasi yang dilakukan dengan human evaluation, dimana dilakukan validasi dari ahli mengenai akurasi, relevansi dan juga tingkat halusinasi. Dari tabel 8 menunjukkan bahwa:

a. Akurasi mengalami peningkatan di semua model yaitu Llama3.1:8b dari 1.61 menjadi 4.34, Llama3.2:1b dari 1.81 menjadi 3.82 sedangkan model Llama3.2:3b juga mengalami peningkatan yang signifikan yaitu dari 1.88 menjadi 4.33.
b. Relevansi juga mengalami peningkatan di semua model yaitu Llama3.1:8b dari 1.92 menjadi 4.59, Llama3.2:1b dari 2.10 menjadi 4.18 sedangkan model Llama3.2:3b mengalami peningkatan dari 2.18 menjadi 4.58.
c. Halusinasi mengalami penurunan yaitu model Llama3.1:8b dari 4.00 menjadi 1.37, Llama3.2:1b dari 3.98 menjadi 1.98 sedangkan model Llama3.2:3b dari 3.95 menjadi 1.41.

4. KESIMPULAN

Penelitian ini membuktikan bahwa metode RAG secara signifikan mampu meningkatkan kualitas jawaban dari berbagai model LLM, baik model kecil maupun menengah. Penerapan RAG memberikan konteks eksternal yang relevan melalui proses retrieval dokumen, sehingga menghasilkan jawaban yang lebih akurat dan sesuai dengan pertanyaan. Hal ini ditunjukkan oleh peningkatan skor F1 secara konsisten pada seluruh metrik ROUGE (ROUGE-1, ROUGE-2, dan ROUGE-L) di semua model, dengan peningkatan tertinggi terjadi pada model Llama3.1:8b. Tidak hanya secara metrik otomatis, hasil ini juga diperkuat melalui evaluasi manual oleh manusia (human evaluation) yang menunjukkan peningkatan yang selaras. Seluruh model mengalami peningkatan dalam aspek

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2120−2128
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.8048

akurasi dan relevansi, serta penurunan yang signifikan pada aspek halusinasi. Sebagai contoh, model Llama3.1:8b mengalami peningkatan akurasi dari 1.61 menjadi 4.34 dan penurunan halusinasi dari 4.00 menjadi 1.37. Hal ini menunjukkan bahwa keberadaan konteks dokumen sangat penting dalam menekan kesalahan semantik dan informasi yang tidak faktual. Secara keseluruhan, dapat disimpulkan bahwa efektivitas RAG tidak hanya bergantung pada ukuran model, tetapi lebih pada kemampuan sistem dalam memanfaatkan informasi eksternal yang relevan. Oleh karena itu, metode RAG sangat direkomendasikan untuk diintegrasikan dalam sistem tanya jawab berbasis dokumen, terutama pada penggunaan model LLM lokal dengan sumber daya terbatas. Integrasi ini terbukti dapat meningkatkan akurasi semantik, mengurangi kesalahan halusinatif, dan menghasilkan jawaban yang lebih dapat dipercaya serta bermanfaat bagi pengguna akhir.

REFERENCES

[1] A. Plaat, M. van Duijn, N. van Stein, M. Preuss, P. van der Putten, and K. J. Batenburg, “Agentic Large Language Models, a survey,” 2025, [Online]. Available: http://arxiv.org/abs/2503.23037
[2] L. Huang et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–58, 2025, doi: 10.1145/3703155.
[3] Wikipedia, “Retrieval-Augmented Generation,” 2024. [Online]. Available: https://en.wikipedia.org/wiki/Retrieval-augmented_generation
[4] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” pp. 1–21, 2023, [Online]. Available: http://arxiv.org/abs/2312.10997
[5] Y. Ding et al., “A Survey on RAG Meets LLMs: Towards Retrieval-Augmented Large Language Models,” 2024, [Online]. Available: http://arxiv.org/abs/2405.06211
[6] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Adv. Neural Inf. Process. Syst., vol. 2020-Decem, 2020.
[7] V. Karpukhin et al., “Dense passage retrieval for open-domain question answering,” EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.
[8] Z. Zhang, Y. Feng, and M. Zhang, “LevelRAG: Enhancing Retrieval-Augmented Generation with Multi-hop Logic Planning over Rewriting Augmented Searchers,” 2025, [Online]. Available: http://arxiv.org/abs/2502.18139
[9] F. Liu, Z. Kang, and X. Han, “Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models Optimizing RAG Techniques Based on Locally Deployed Ollama Models A Case Study with Locally Deployed Ollama Models,” 2024, [Online]. Available: https://arxiv.org/abs/2408.05933
[10] G. F. Febrian and G. Figueredo, “KemenkeuGPT: Leveraging a Large Language Model on Indonesia’s Government Financial Data and Regulations to Enhance Decision Making,” 2024, [Online]. Available: http://arxiv.org/abs/2407.21459
[11] A. F. Aji et al., “One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia,” Proc. Annu. Meet. Assoc. Comput. Linguist., vol. 1, pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.
[12] N. Livathinos et al., “Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion,” 2025, [Online]. Available: http://arxiv.org/abs/2501.17887
[13] A.-L. Bornea, F. Ayed, A. De Domenico, N. Piovesan, and A. Maatouk, “Telco-RAG: Navigating the Challenges of Retrieval-Augmented Language Models for Telecommunications,” 2024, [Online]. Available: http://arxiv.org/abs/2404.15939
[14] The PostgreSQL Global Development Group, “pgvector 0.7.0 Released!” [Online]. Available: https://www.postgresql.org/about/news/pgvector-070-released-2852/
[15] J. Jenq, “Improving Performance of Local Chatbot with Caching,” Proc. 28th World Multi-Conference Syst. Cybern. Informatics WMSCI 2024, vol. 22, no. 5, pp. 68–71, 2024, doi: 10.54808/wmsci2024.01.68.
[16] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” 2023, [Online]. Available: http://arxiv.org/abs/2302.13971
[17] T. Rehman, S. Ghosh, K. Das, S. Bhattacharjee, D. K. Sanyal, and S. Chattopadhyay, “Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets,” 2025, [Online]. Available: https://huggingface.co/unsloth/llama-3-8b-bnb-4bit
[18] H. Nguyen, H. Chen, L. Pobbathi, and J. Ding, “A Comparative Study of Quality Evaluation Methods for Text Summarization,” 2024, [Online]. Available: http://arxiv.org/abs/2407.00747
[19] E. Kamalloo, N. Dziri, C. L. A. Clarke, and D. Rafiei, “Evaluating Open-Domain Question Answering in the Era of Large Language Models,” Proc. Annu. Meet. Assoc. Comput. Linguist., vol. 1, pp. 5591–5606, 2023, doi: 10.18653/v1/2023.acl-long.307.
[20] A. Joshi, S. Kale, S. Chandel, and D. Pal, “Likert Scale: Explored and Explained,” Br. J. Appl. Sci. Technol., vol. 7, no. 4, pp. 396–403, 2015, doi: 10.9734/bjast/2015/14975.
[21] LangChain, “How to recursively split text by characters,” 2025. [Online]. Available: https://python.langchain.com/docs/how_to/recursive_text_splitter/
[22] L. Caspari, K. G. Dastidar, S. Zerhoudi, J. Mitrovic, and M. Granitzer, “Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems,” CEUR Workshop Proc., vol. 3784, pp. 62–70, 2024.
[23] J. Isbarov and K. Huseynova, “Enhanced document retrieval with topic embeddings,” 18th IEEE Int. Conf. Appl. Inf. Commun. Technol. AICT 2024, 2024, doi: 10.1109/AICT61888.2024.10740455.

This Journal is licensed under a Creative Commons Attribution 4.0 International License