

Pengembangan Aplikasi *Question Answering* Berbasis *Retrieval-Augmented Generation* Menggunakan Qwen3-8B untuk Pencarian Informasi Toko Elektronik di Palembang

Dhini Novely Saputri ^{1*}, Heri Suroyo ², Leon A. Abdillah ³, M. Soekarno Putra ⁴

^{1*,2,3,4} Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Bina Darma, Kota Palembang, Provinsi Sumatera Selatan, Indonesia.

Corresponding Email: dhininovely19@gmail.com ^{1*}

Histori Artikel:

Dikirim 10 September 2026; *Diterima dalam bentuk revisi* 14 September 2026; *Diterima* 20 September 2026; *Diterbitkan* 30 September 2026. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STM IK Indonesia Banda Aceh.

Abstrak

Large Language Model (LLM) mampu menghasilkan respons bahasa alami, tetapi pada domain pengetahuan spesifik masih berisiko memberikan jawaban yang tidak didukung data. Penelitian ini mengembangkan dan mengevaluasi aplikasi question answering untuk pencarian informasi toko elektronik di Kota Palembang dengan mengintegrasikan Qwen3-8B dan Retrieval-Augmented Generation (RAG). Penelitian menggunakan Design Science Research dengan artefak berupa aplikasi web. Basis pengetahuan memuat 1.019 data produk dari delapan toko elektronik dengan enam atribut: nama toko, alamat, nomor telepon, tautan Google Maps, kategori, dan nama produk. Data diubah menjadi embedding 768 dimensi menggunakan sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 dan disimpan dalam ChromaDB. Sistem mengambil lima data teratas berdasarkan cosine similarity dan menerapkan ambang batas 0,65 sebelum konteks diberikan kepada Qwen3-8B melalui Groq API. Evaluasi menggunakan 200 kueri; label positif menunjukkan informasi tersedia dalam basis pengetahuan, sedangkan label negatif menunjukkan informasi tidak tersedia atau di luar cakupan. Respons dinilai sebagai true positive (TP) jika kueri positif dijawab benar, true negative (TN) jika kueri negatif ditolak dengan tepat, false positive (FP) jika kueri negatif diberi jawaban, dan false negative (FN) jika kueri positif ditolak atau gagal dijawab. Pengujian menghasilkan TP=143, TN=5, FP=34, dan FN=18, dengan accuracy 74,00%, precision 80,79%, recall 88,82%, dan F1-score 84,61%. Kesalahan ditemukan pada kueri pendek yang ambigu, kemiripan semantik, variasi istilah lokal, dan salah ketik. RAG menyediakan konteks retrieval sebagai dasar jawaban, tetapi kemampuan menolak kueri di luar basis pengetahuan masih perlu ditingkatkan.

Kata Kunci: Question Answering; Retrieval-Augmented Generation; Qwen3-8B; ChromaDB; Semantic Search.

Abstract

Large Language Models (LLMs) can generate fluent responses but may produce unsupported answers in domain-specific settings. This study developed and evaluated a question answering application for electronic-store information retrieval in Palembang by integrating Qwen3-8B with Retrieval-Augmented Generation (RAG). Using Design Science Research, the study produced a web application supported by a knowledge base of 1,019 product records from eight stores. Each record contained six attributes: store name, address, telephone number, Google Maps link, category, and product name. The data were converted into 768-dimensional embeddings using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 and stored in ChromaDB. The system retrieved the top five records based on cosine similarity and applied a 0.65 threshold before supplying context to Qwen3-8B through the Groq API. Evaluation used 200 queries; positive labels indicated available information, whereas negative labels indicated unavailable or out-of-scope information. Responses were classified as true positives (TP) when positive queries were answered correctly, true negatives (TN) when negative queries were rejected correctly, false positives (FP) when negative queries received answers, and false negatives (FN) when positive queries were rejected or not answered. Testing yielded TP=143, TN=5, FP=34, and FN=18, with 74.00% accuracy, 80.79% precision, 88.82% recall, and an 84.61% F1-score. Errors occurred with short ambiguous queries, semantic overlap, local terminology, and typographical variation. RAG supplied retrieved context as the basis for responses, but rejection of queries outside the knowledge base remains limited.

Keyword: Question Answering; Retrieval-Augmented Generation; Qwen3-8B; ChromaDB; Semantic Search.

1. Pendahuluan

Perkembangan *Large Language Model* (LLM) memperluas kemampuan sistem pemrosesan bahasa alami dalam memahami instruksi dan menghasilkan respons yang koheren. Namun, model generatif masih berisiko menghasilkan halusinasi, yaitu informasi yang tampak meyakinkan tetapi tidak didukung fakta atau sumber pengetahuan yang valid. Risiko ini penting pada aplikasi pencarian informasi dalam domain tertentu karena pengguna membutuhkan jawaban yang dapat ditelusuri ke data, bukan hanya mengandalkan pengetahuan parametrik model (Huang *et al.*, 2025).

Retrieval-Augmented Generation (RAG) memadukan pengambilan informasi dari sumber pengetahuan eksternal dengan pembentukan jawaban oleh LLM. Dalam mekanisme ini, kueri pengguna digunakan untuk mengambil konteks yang relevan, lalu konteks tersebut diberikan kepada model generatif sebagai dasar jawaban. Lewis *et al.* (2020) menunjukkan manfaat akses terhadap memori nonparametrik pada tugas yang membutuhkan pengetahuan faktual. Kualitas RAG juga bergantung pada ketepatan pengambilan konteks, kemampuan model menggunakan konteks tersebut, dan pengendalian respons ketika konteks yang diperoleh tidak relevan (Gao *et al.*, 2023; Chen *et al.*, 2024).

Penggunaan RAG tidak secara otomatis menghilangkan halusinasi. RAGTruth menunjukkan bahwa respons berbasis RAG masih dapat memuat klaim yang tidak didukung atau bertentangan dengan konteks yang diambil (Niu *et al.*, 2024). Oleh karena itu, evaluasi RAG perlu membedakan kemampuan menemukan konteks, ketepatan jawaban, dan kemampuan menolak kueri di luar basis pengetahuan. ARES menilai relevansi konteks, kesetiaan jawaban terhadap konteks (*faithfulness*), dan relevansi jawaban (Saad-Falcon *et al.*, 2024), sedangkan RAGAS menawarkan evaluasi terpisah terhadap pengambilan konteks dan pembentukan jawaban (Es *et al.*, 2023). Untuk mendukung pencarian semantik, Sentence-BERT memperkenalkan representasi kalimat berbasis *embedding* yang dapat dibandingkan menggunakan kemiripan kosinus (Reimers & Gurevych, 2019). Laporan teknis Qwen3 juga menjelaskan dukungan multibahasa pada model tersebut (Yang *et al.*, 2025).

Di Kota Palembang, kebutuhan penelitian diidentifikasi melalui observasi dan dokumentasi pada delapan toko elektronik yang menjadi objek studi. Informasi yang dibutuhkan calon konsumen—nama produk, kategori, nama toko, alamat, nomor telepon, dan tautan Google Maps—berada pada sumber atau toko yang terpisah. Akibatnya, pencarian satu produk dapat memerlukan penelusuran beberapa sumber secara manual. Delapan toko tersebut merupakan cakupan studi kasus berdasarkan objek yang berhasil diobservasi dan didokumentasikan, bukan sampel yang diklaim mewakili seluruh toko elektronik di Palembang. Temuan ini menjadi dasar penyusunan basis pengetahuan terstruktur agar pengguna dapat menanyakan produk dan memperoleh informasi toko terkait melalui satu antarmuka.

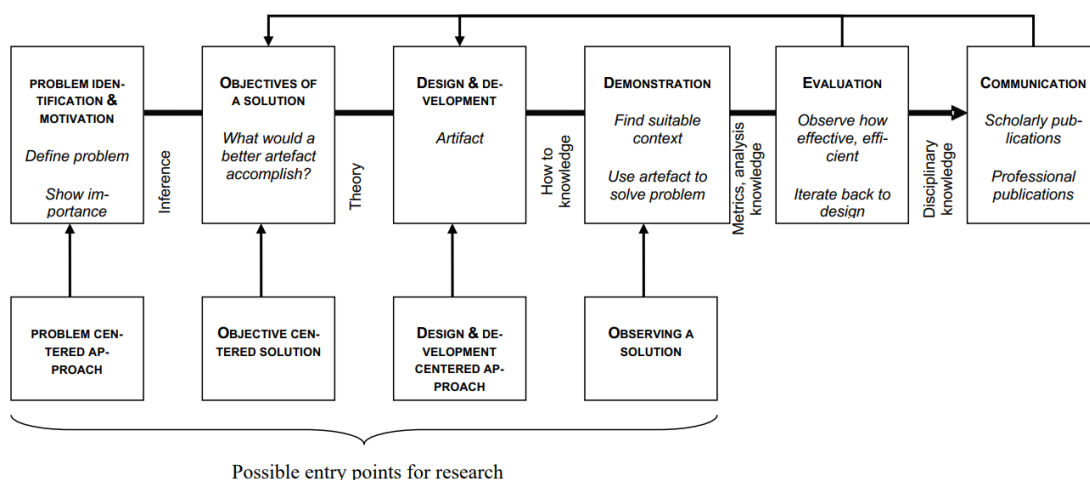
Penelitian RAG mencakup pengolahan korpus dokumen umum, pengujian pada tolok ukur lintas domain, serta pengembangan mekanisme pengambilan konteks adaptif. Self-RAG mengembangkan pengambilan konteks adaptif disertai mekanisme refleksi diri (Asai *et al.*, 2024), sedangkan Corrective RAG menambahkan evaluasi hasil pengambilan konteks untuk mengurangi dampak dokumen yang tidak relevan (Yan *et al.*, 2024). Penelitian ini berfokus pada penerapan RAG pada basis pengetahuan terstruktur berskala lokal yang memuat relasi produk–toko dan informasi navigasi. Evaluasinya mencakup pencarian informasi yang tersedia serta penolakan permintaan yang tidak didukung basis pengetahuan, termasuk kebutuhan untuk mengembalikan atribut toko yang tepat.

Berdasarkan kebutuhan tersebut, penelitian ini bertujuan mengembangkan aplikasi *question answering* berbasis web menggunakan Qwen3-8B, ChromaDB, dan *embedding* multibahasa untuk pencarian informasi toko elektronik di Palembang. Kinerja aplikasi dievaluasi menggunakan matriks konfusi, akurasi, presisi, *recall*, dan skor F1 pada 200 kueri uji. Kontribusi penelitian mencakup penerapan alur RAG pada data produk–toko terstruktur, perancangan aturan agar jawaban mengacu pada konteks hasil pengambilan informasi, serta analisis kasus positif palsu (*false positive*) dan negatif palsu (*false negative*) pada kueri bahasa alami.

2. Metode Penelitian

Penelitian ini menggunakan *Design Science Research Methodology* (DSRM) karena berorientasi pada pembangunan dan evaluasi artefak untuk menyelesaikan masalah praktis. Tahapan DSRM meliputi identifikasi masalah dan motivasi, penentuan tujuan solusi, perancangan dan pengembangan, demonstrasi, evaluasi, serta komunikasi (Peppers *et al.*, 2007). Artefak yang dikembangkan dalam penelitian ini adalah aplikasi *question answering* berbasis web untuk pencarian informasi toko elektronik. Tahapan penelitian mengacu pada model proses yang ditunjukkan pada Gambar 1.

Nominal process sequence



Gambar 1. Model proses *Design Science Research Methodology* (DSRM) yang digunakan dalam penelitian.

2.1 Dataset dan Basis Pengetahuan

Data primer diperoleh melalui observasi dan dokumentasi pada delapan toko elektronik di Kota Palembang. Informasi yang dicatat meliputi nama toko, alamat, nomor telepon, tautan Google Maps, kategori, dan nama produk. Hubungan produk–toko diverifikasi berdasarkan hasil observasi dan dokumentasi masing-masing toko sebelum data diseragamkan. Prapemrosesan mencakup pemeriksaan kelengkapan dan konsistensi, pembersihan, penghapusan duplikat, serta normalisasi format penulisan. Kumpulan data akhir memuat 1.019 baris; setiap baris merepresentasikan hubungan produk dan toko melalui enam atribut pada Tabel 1. Data tersebut digunakan sebagai basis pengetahuan statis dan tidak memuat harga maupun jumlah stok waktu nyata. Periode observasi tidak tercatat secara spesifik dalam dokumentasi penelitian.

Tabel 1. Struktur kumpulan data penelitian

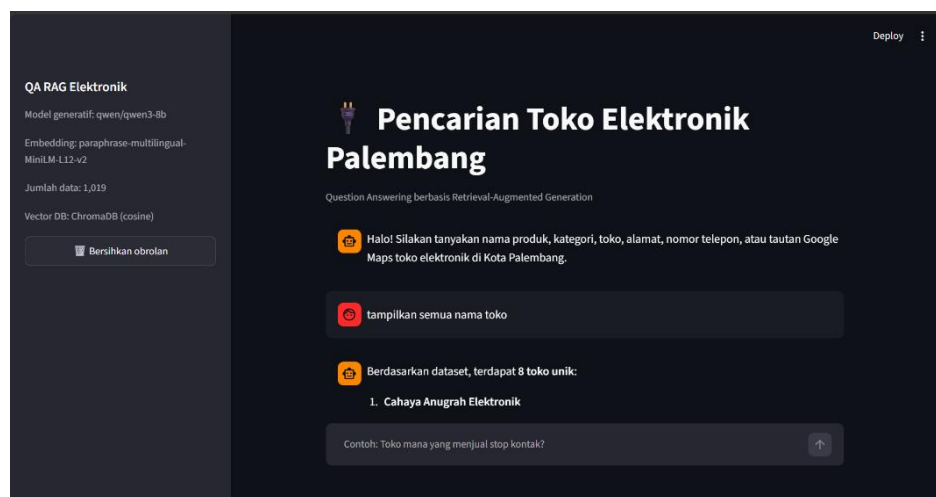
Atribut	Keterangan
Nama_Toko	Nama toko elektronik di Kota Palembang
Alamat	Alamat lengkap toko
No_HP	Nomor telepon atau kontak toko
Link_Map	Tautan Google Maps lokasi toko
Kategori	Kategori produk elektronik
Nama_Produk	Nama produk elektronik yang tercatat

2.2 Arsitektur RAG dan Implementasi

Setiap baris data dibentuk menjadi teks dokumen, kemudian diubah menjadi *embedding* berdimensi 768 menggunakan model sentence-transformers/paraphrase-multilingual-MiniLM-L12-

v2. Representasi vektor beserta metadata Nama_Produk, Kategori, Nama_Toko, Alamat, No_HP, dan Link_Map disimpan dalam ChromaDB. Kueri pengguna diubah menjadi *embedding* menggunakan model yang sama, lalu dibandingkan dengan data tersimpan berdasarkan *cosine similarity*. Sistem mengambil lima data dengan kemiripan tertinggi ($top-k = 5$) dan menerapkan ambang batas 0,65 untuk menentukan relevansi hasil pengambilan. Nilai $top-k$ dan ambang batas tersebut digunakan secara tetap pada pengujian 200 kueri. Penelitian ini belum menguji pengaruh variasi kedua parameter terhadap kinerja sistem.

Konteks hasil pengambilan digabungkan dengan pertanyaan pengguna dalam *system prompt* yang memuat aturan pembatasan jawaban. Model diarahkan untuk menjawab hanya berdasarkan konteks yang diberikan dan menggunakan atribut produk, kategori, nama toko, alamat, nomor telepon, serta tautan Google Maps yang tersedia. Model juga diarahkan untuk tidak membuat atau menghitung harga dan jumlah stok waktu nyata karena kedua atribut tersebut tidak terdapat dalam basis pengetahuan. Apabila informasi tidak ditemukan, tidak relevan, atau berada di luar cakupan data, model diarahkan untuk menolak secara sopan dan menyatakan bahwa informasi tidak tersedia. *Prompt* dikirim ke Qwen3-8B melalui Groq API, kemudian respons ditampilkan melalui Streamlit dalam bentuk kartu informasi. Pencarian menggunakan *embedding* semantik dan kemiripan kosinus (Reimers & Gurevych, 2019), sedangkan Qwen3-8B digunakan sebagai model pembentuk jawaban multibahasa (Yang *et al.*, 2025). Tampilan utama aplikasi ditunjukkan pada Gambar 2.



Gambar 2. Tampilan utama aplikasi *question answering* berbasis RAG.

2.3 Skenario dan Metrik Evaluasi

Evaluasi sumatif menggunakan 200 kueri yang ditentukan sebelum pengujian. Sebanyak 100 kueri diberi label positif karena informasi yang diminta tersedia dalam basis pengetahuan, sedangkan 100 kueri diberi label negatif karena informasi tidak tersedia atau berada di luar cakupan data. Kueri mencakup tujuh skenario: pencarian nama toko berdasarkan produk, kategori produk, nama produk, alamat toko, nomor telepon, lokasi atau tautan Google Maps, serta informasi di luar kumpulan data. Respons sistem dibandingkan secara manual dengan data acuan. Dokumentasi penelitian tidak mencatat pemisahan peran antara pembuat kueri dan penilai respons; karena itu, penelitian ini tidak mengklaim adanya evaluator independen atau pengukuran kesepakatan antarpenilai. Distribusi kueri disajikan pada Tabel 2.

Tabel 2. Distribusi 200 kueri pengujian

Skenario	Positif	Negatif	Total
Nama toko berdasarkan produk	14	20	34
Kategori produk	13	15	28
Nama produk	14	15	29

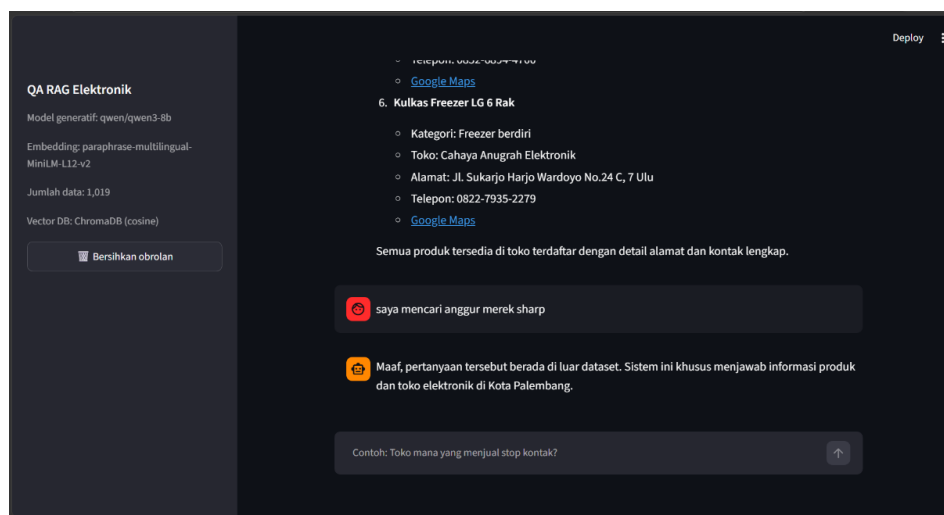
Alamat toko	14	15	29
Nomor telepon toko	14	15	29
Lokasi/Google Maps	11	15	26
Informasi di luar kumpulan data	20	5	25
Jumlah	100	100	200

Penilaian respons didasarkan pada ketersediaan informasi dalam basis pengetahuan. *True positive* (TP) adalah kueri positif yang dijawab sesuai data; *true negative* (TN) adalah kueri negatif yang ditolak dengan tepat; *false positive* (FP) adalah kueri negatif yang diberi jawaban seolah-olah didukung data; dan *false negative* (FN) adalah kueri positif yang ditolak atau gagal dijawab. Untuk pertanyaan tentang alamat, nomor telepon, dan tautan Google Maps, jawaban dinilai benar apabila atributnya sesuai dengan toko atau produk terkait dalam kumpulan data. Akurasi dihitung sebagai $(TP + TN)/(TP + TN + FP + FN)$, presisi sebagai $TP/(TP + FP)$, *recall* sebagai $TP/(TP + FN)$, dan skor F1 sebagai rata-rata harmonis presisi dan *recall*.

3. Hasil dan Pembahasan

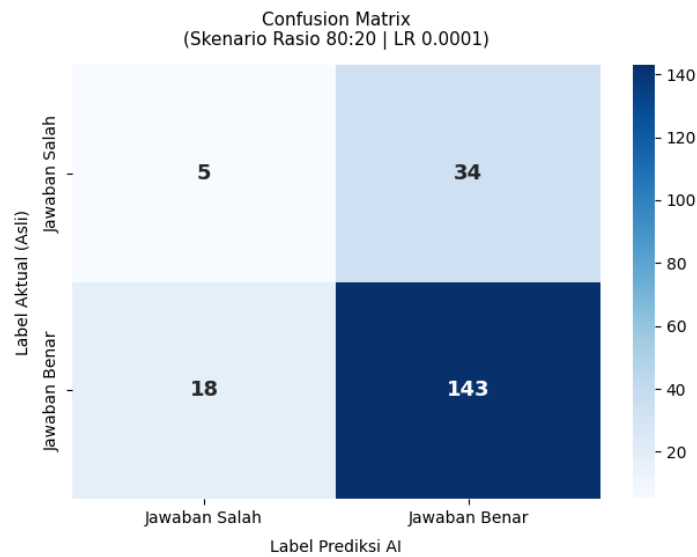
3.1 Hasil

Aplikasi berhasil menerima kueri bahasa alami dan mengembalikan informasi toko berdasarkan hasil pengambilan data dari ChromaDB. Pada kueri yang didukung basis pengetahuan, keluaran dapat memuat nama produk, kategori, nama toko, alamat, nomor kontak, dan tautan Google Maps. Pada kueri non-elektronik atau permintaan atribut yang tidak tersedia, aturan pembatasan jawaban dirancang untuk memberikan pesan bahwa informasi tidak tersedia tanpa membuat detail tambahan. Contoh hasil pengambilan data dan penyajian kartu toko ditunjukkan pada Gambar 3.



Gambar 3. Contoh hasil pencarian informasi toko berdasarkan kueri pengguna.

Dari 200 kueri, sistem menghasilkan 143 *true positive* (TP), 5 *true negative* (TN), 34 *false positive* (FP), dan 18 *false negative* (FN), sehingga terdapat 148 respons yang dinilai benar dan 52 respons yang dinilai salah. Nilai $TN=5$ menunjukkan bahwa penolakan kueri negatif masih menjadi kelemahan sistem. Visualisasi matriks konfusi ditunjukkan pada Gambar 4, sedangkan hasil perhitungan metrik disajikan pada Tabel 3.



Gambar 4. Matriks konfusi hasil pengujian 200 kueri.

Tabel 3. Hasil evaluasi kinerja sistem

Metrik	Nilai
Akurasi	74,00%
Presisi	80,79%
Recall	88,82%
Skor F1	84,61%

Akurasi 74,00% menunjukkan bahwa 148 dari 200 respons dinilai sesuai dengan kondisi aktual. Presisi 80,79% menunjukkan bahwa sebagian besar jawaban positif yang diberikan sistem dinilai benar, meskipun 34 *false positive* memperlihatkan bahwa sistem masih memberikan jawaban pada sejumlah kueri negatif. *Recall* 88,82% merupakan nilai metrik tertinggi dan menunjukkan bahwa sistem berhasil menjawab sebagian besar kueri yang informasinya tersedia. Skor F1 sebesar 84,61% merangkum presisi dan *recall* melalui rata-rata harmonis keduanya. Distribusi kueri menurut skenario telah disajikan pada Tabel 2. Namun, rekap TP, TN, FP, dan FN untuk masing-masing dari tujuh skenario serta nilai *cosine similarity* setiap kueri tidak tersimpan dalam dokumentasi hasil pengujian. Karena itu, angka hasil per skenario dan nilai kemiripan per kueri tidak dapat dilaporkan. Penelusuran kesalahan dalam penelitian ini didasarkan pada contoh kasus yang terdokumentasi dan parameter pengambilan data yang digunakan. Contoh kueri negatif meliputi “Di mana saya bisa mendapatkan anggur merek Sharp 2 pintu?”, “Toko mana yang menjual PlayStation 6 di Palembang?”, serta pertanyaan tentang harga atau layanan yang atributnya tidak tersedia. Kueri negatif dinilai benar hanya jika sistem menolak memberikan informasi yang tidak didukung basis pengetahuan.

3.2 Pembahasan

Sebanyak 34 *false positive* ditemukan dalam pengujian. Catatan kesalahan menunjukkan bahwa sebagian kasus berkaitan dengan kueri pendek yang ambigu atau memiliki tumpang tindih semantik dengan entri basis pengetahuan. Frasa seperti “AC” atau “HP”, misalnya, dapat menarik entri dengan kemiripan semantik meskipun maksud pengguna berbeda. Nilai kemiripan untuk masing-masing kasus tidak direkam sehingga hubungan tersebut tidak dapat dijelaskan secara numerik. Temuan ini berkaitan dengan pembahasan Chen *et al.* (2024) mengenai tantangan penolakan kueri negatif dan ketahanan RAG terhadap konteks yang tidak tepat: ketika pengambilan data keliru, model generatif dapat mengikuti konteks yang salah.

Sebanyak 18 *false negative* juga tercatat dalam pengujian. Variasi istilah lokal, bentuk tidak baku, dan salah ketik diduga berkaitan dengan kegagalan sistem menjawab sebagian kueri yang informasinya tersedia. Dalam implementasi penelitian, ambang batas 0,65 digunakan untuk menentukan relevansi hasil pengambilan data. Karena pengujian hanya menggunakan satu konfigurasi ambang batas dan $top-k = 5$, penelitian ini belum dapat menyimpulkan bahwa kedua nilai tersebut optimal. Pengujian beberapa nilai ambang batas dan $top-k$ diperlukan untuk mengukur hubungan antara kemampuan menemukan data dan kemampuan menolak konteks yang tidak relevan.

Hasil penelitian perlu ditafsirkan secara hati-hati terkait halusinasi. Presisi, *recall*, dan skor F1 tidak secara langsung mengukur *faithfulness* atau halusinasi pada tingkat klaim. RAGTruth menunjukkan bahwa model masih dapat menghasilkan pernyataan yang tidak sepenuhnya didukung konteks meskipun menggunakan RAG (Niu *et al.*, 2024), sedangkan ARES mengusulkan pengukuran *faithfulness* secara khusus (Saad-Falcon *et al.*, 2024). Penelitian ini menggunakan matriks konfusi berdasarkan kesesuaian respons dengan basis pengetahuan. Oleh karena itu, hasilnya menunjukkan kinerja sistem dalam menjawab atau menolak kueri menurut aturan penilaian yang digunakan, tetapi belum membuktikan bahwa halusinasi telah dihilangkan atau berkurang dibandingkan sistem tanpa RAG.

Dibandingkan pendekatan RAG yang mengambil dokumen bebas, penelitian ini menggunakan data produk–toko terstruktur sehingga konteks hasil pengambilan dapat dipetakan ke atribut yang dibutuhkan pengguna. Namun, pengambilan lima dokumen secara tetap dapat menyertakan konteks yang kurang relevan pada kueri ambigu. Self-RAG menunjukkan penggunaan pengambilan konteks adaptif untuk mengurangi pengambilan konteks yang tidak diperlukan (Asai *et al.*, 2024). Pendekatan tersebut belum diuji dalam penelitian ini. Keterbatasan evaluasi meliputi tidak tersedianya catatan nilai kemiripan per kueri, rekap matriks konfusi per skenario, pengujian sensitivitas ambang batas dan $top-k$, serta pengukuran waktu inferensi Groq. Waktu pemrosesan dari awal hingga akhir dan waktu inferensi menurut panjang kueri juga tidak direkam pada eksperimen asli, sehingga tidak dapat dihitung kembali dari dokumentasi yang tersedia. Penelitian lanjutan perlu mencatat metrik waktu tersebut dan menguji pencarian hibrida atau pemeringkatan ulang, normalisasi istilah lokal, relevansi konteks, serta *faithfulness* jawaban.

4. Kesimpulan

Penelitian menghasilkan aplikasi *Question Answering* berbasis *Retrieval-Augmented Generation* menggunakan Qwen3-8B untuk pencarian informasi toko elektronik di Kota Palembang. Artefak mengintegrasikan *knowledge base* berisi 1.019 data produk dari delapan toko, *embedding multilingual*, ChromaDB, *grounding rules*, dan Qwen3-8B melalui Groq API. Evaluasi menunjukkan bahwa sistem lebih kuat dalam menemukan informasi yang tersedia dibandingkan menolak kueri yang tidak didukung data; *false positive* terutama berkaitan dengan kueri pendek/ambigu, sedangkan *false negative* berkaitan dengan variasi istilah dan salah ketik. Temuan harus diposisikan sebagai pencarian terhadap *knowledge base* statis pada saat data dikumpulkan, bukan sebagai informasi ketersediaan produk terkini, karena *dataset* tidak memuat harga maupun stok *real-time* dan tidak terhubung langsung dengan inventori toko mitra. Pengembangan berikutnya perlu menguji sensitivitas *threshold* dan $top-k$, menerapkan *hybrid search* atau *reranking*, memperkaya sinonim lokal, menghubungkan data dinamis melalui API toko, serta menambahkan evaluasi khusus *context relevance*, *faithfulness*, *hallucination*, dan *latency*.

5. Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada Universitas Bina Darma dan toko-toko elektronik di Palembang yang telah mendukung pelaksanaan penelitian dan pengumpulan data. Penulis juga

berterima kasih kepada dosen pembimbing atas bimbingan, arahan, dan dukungannya selama penelitian hingga penyusunan artikel ini.

6. Daftar Pustaka

- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations (ICLR)*.
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- Gupta, S., Ranjan, R., & Singh, S. N. (2024). *A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2410.12837>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. <https://doi.org/10.1145/3703155>
- Jin, J., Zhu, Y., Yang, X., Zhang, C., & Dou, Z. (2024). *FlashRAG: A modular toolkit for efficient retrieval-augmented generation research* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2405.13576>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10862–10878). <https://doi.org/10.18653/v1/2024.acl-long.585>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). <https://doi.org/10.18653/v1/D19-1410>

- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 338–354). <https://doi.org/10.18653/v1/2024.naacl-long.20>
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., . . . Qiu, Z. (2025). *Qwen3 technical report* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2505.09388>.