



Pengembangan Sistem Informasi Otomotif Berbasis *Chatbot* Menggunakan DeBERTa

Fadhel Putra Felmi^{1*}, Viny Cristanti M²

¹⁻²Teknik Informatika, Universitas Tarumanagara, Indonesia

*Penulis Korespondensi: fadhel.535220182@stu.untar.ac.id

Abstract. This study aims to develop an automotive chatbot system capable of providing contextual, accurate, and reference-based responses using the Retrieval-augmented generation (RAG) approach. Conventional automotive information services often face limitations in delivering relevant technical information quickly and accurately. Large Language Models (LLMs) provide natural responses but still suffer from hallucination issues that may reduce answer reliability. To address this limitation, the proposed system integrates semantic retrieval using Multilingual E5 and reranking using DeBERTa v3 before generating responses through a Large Language Model. User queries and automotive documents are transformed into semantic vector representations, enabling retrieval based on contextual similarity rather than keyword matching. Retrieved documents are subsequently reranked to improve contextual relevance before being used as supporting knowledge during answer Generation. The system is implemented as a web-based application and evaluated using retrieval metrics and the RAGAS framework. Experimental results indicate that the proposed approach improves contextual understanding and answer relevance while reducing the possibility of hallucinated responses. The developed chatbot demonstrates the potential of Retrieval-augmented generation for supporting intelligent automotive information services. The evaluation results show that the proposed system achieved a Faithfulness score of 1.000, an Answer Relevancy score of 0.883, and a Context Precision score of 1.000. These results indicate that the integration of Multilingual E5 and DeBERTa v3 successfully improves contextual relevance while reducing the risk of hallucinated responses.

Keywords: Chatbot; DeBERTa v3; Multilingual e5; Retrieval-Augmented Generation; Semantic Retrieval.

Abstrak. Penelitian ini bertujuan untuk mengembangkan sistem *chatbot* otomotif yang mampu memberikan jawaban kontekstual, akurat, dan berbasis referensi menggunakan pendekatan *Retrieval-augmented generation* (RAG). Layanan informasi otomotif konvensional masih memiliki keterbatasan dalam menyediakan informasi teknis yang cepat dan relevan kepada pengguna. Di sisi lain, penggunaan Large Language Model (LLM) mampu menghasilkan jawaban yang natural, namun masih berpotensi menghasilkan informasi yang tidak sesuai fakta atau dikenal sebagai hallucination. Untuk mengatasi permasalahan tersebut, penelitian ini mengintegrasikan *semantic retrieval* menggunakan Multilingual E5 dan reranking menggunakan DeBERTa v3 sebelum proses generasi jawaban dilakukan oleh Large Language Model. Pertanyaan pengguna dan dokumen otomotif diubah menjadi representasi vektor semantik sehingga proses pencarian dapat dilakukan berdasarkan kesamaan makna. Dokumen hasil *retrieval* kemudian diurutkan kembali menggunakan DeBERTa v3 untuk meningkatkan relevansi konteks yang digunakan dalam proses pembangkitan jawaban. Sistem diimplementasikan dalam bentuk aplikasi berbasis web dan dievaluasi menggunakan metrik *retrieval* serta framework RAGAS. Hasil penelitian menunjukkan bahwa pendekatan yang diusulkan mampu meningkatkan relevansi jawaban, memahami konteks pertanyaan dengan lebih baik, serta membantu mengurangi potensi hallucination pada *chatbot*. Dengan demikian, sistem yang dikembangkan dapat dimanfaatkan sebagai layanan informasi otomotif berbasis kecerdasan buatan yang lebih efektif. Hasil evaluasi menunjukkan bahwa sistem memperoleh nilai Faithfulness sebesar 1,000, Answer Relevancy sebesar 0,883, dan Context Precision sebesar 1,000. Hasil tersebut menunjukkan bahwa integrasi Multilingual E5 dan DeBERTa v3 mampu meningkatkan relevansi jawaban serta membantu mengurangi potensi hallucination pada *chatbot*.

Kata kunci: Chatbot; DeBERTa v3; Multilingual E5; *Retrieval-Augmented Generation*; Temu Kembali Semantik.

1. LATAR BELAKANG

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence/AI*) telah mendorong peningkatan kemampuan sistem komputer dalam memahami dan menghasilkan bahasa alami. Salah satu implementasi yang berkembang pesat adalah *chatbot*, yaitu sistem yang mampu berinteraksi dengan pengguna melalui percakapan berbasis teks maupun suara.

Chatbot telah banyak dimanfaatkan pada berbagai bidang, seperti pendidikan, kesehatan, layanan pelanggan, dan sistem informasi karena mampu menyediakan informasi secara cepat, konsisten, dan tersedia selama dua puluh empat jam (Elysia & Herianto, 2024; Sianipar et al., 2022).

Dalam bidang otomotif, kebutuhan terhadap layanan informasi yang cepat dan akurat terus meningkat seiring bertambahnya jumlah pengguna kendaraan bermotor. Pengguna sering menghadapi berbagai permasalahan teknis, seperti gangguan mesin, sistem kelistrikan, sistem pengereman, sistem pendingin, maupun perawatan kendaraan. Informasi mengenai permasalahan tersebut umumnya diperoleh melalui pencarian internet, forum diskusi, atau konsultasi langsung dengan teknisi. Namun, proses tersebut sering kali membutuhkan waktu yang relatif lama dan belum tentu menghasilkan informasi yang sesuai dengan kebutuhan pengguna.

Perkembangan *Large Language Model* (LLM) memberikan peluang baru dalam pengembangan *chatbot* yang mampu menghasilkan jawaban lebih natural dan menyerupai komunikasi manusia. Model bahasa besar seperti GPT, LLaMA, dan Qwen mampu memahami konteks pertanyaan serta menghasilkan respons yang lebih baik dibandingkan *chatbot* berbasis aturan (*rule-based chatbot*). Meskipun demikian, LLM masih memiliki keterbatasan berupa hallucination, yaitu kondisi ketika sistem menghasilkan informasi yang terlihat benar tetapi tidak didukung oleh sumber informasi yang valid. Permasalahan ini dapat menurunkan tingkat kepercayaan pengguna terhadap sistem, khususnya pada domain otomotif yang membutuhkan informasi teknis yang akurat (Huang et al., 2025; Karpukhin et al., 2020).

Salah satu pendekatan yang banyak digunakan untuk mengatasi permasalahan tersebut adalah *Retrieval-augmented generation* (RAG). Pendekatan ini menggabungkan kemampuan pencarian dokumen dengan kemampuan generatif model bahasa sehingga jawaban yang dihasilkan dapat didukung oleh dokumen referensi yang relevan. Melalui mekanisme *retrieval*, sistem dapat mengambil informasi dari basis pengetahuan eksternal sebelum proses generasi jawaban dilakukan. Dengan demikian, jawaban yang dihasilkan menjadi lebih akurat, kontekstual, dan dapat dipertanggungjawabkan (Hoshi et al., 2023).

Beberapa penelitian sebelumnya menunjukkan bahwa *Retrieval-augmented generation* mampu meningkatkan kualitas jawaban *chatbot* pada berbagai domain. Penerapan RAG pada layanan informasi pendidikan berhasil meningkatkan relevansi jawaban dengan memanfaatkan dokumen sekolah sebagai sumber pengetahuan. Selain itu, penelitian pada bidang kesehatan juga menunjukkan bahwa pendekatan RAG mampu mengurangi risiko hallucination serta meningkatkan kualitas respons yang diberikan kepada pengguna. Meskipun demikian,

implementasi RAG pada domain otomotif masih relatif terbatas, khususnya yang mengombinasikan *semantic retrieval* menggunakan Multilingual E5 dan reranking menggunakan DeBERTa v3 (Elysia & Herianto, 2024; Zhou et al., 2024; Samudra et al., 2025).

Multilingual E5 merupakan model embedding yang mampu menghasilkan representasi semantik berkualitas tinggi untuk berbagai bahasa. Model ini memungkinkan sistem melakukan pencarian dokumen berdasarkan kesamaan makna, bukan hanya kesamaan kata. Sementara itu, DeBERTa v3 merupakan model transformer yang memiliki kemampuan baik dalam mengevaluasi relevansi dokumen terhadap pertanyaan pengguna. Kombinasi kedua model tersebut berpotensi meningkatkan kualitas *retrieval* sehingga konteks yang digunakan pada proses generasi jawaban menjadi lebih relevan (Wang et al., 2024; Gao & Chen, 2023).

Berdasarkan permasalahan tersebut, penelitian ini mengembangkan *chatbot* otomotif berbasis *Retrieval-augmented generation* dengan mengintegrasikan Multilingual E5 sebagai model embedding, FAISS sebagai vector database, DeBERTa v3 sebagai model reranking, dan Qwen2.5 sebagai Large Language Model. Penelitian ini bertujuan menghasilkan sistem *chatbot* yang mampu memberikan jawaban yang lebih relevan, akurat, dan berbasis referensi sehingga dapat membantu pengguna memperoleh informasi otomotif secara lebih efektif. Selain itu, penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan layanan informasi otomotif berbasis kecerdasan buatan yang lebih andal dan terpercaya. Kebaruan penelitian ini terletak pada integrasi Multilingual E5 sebagai semantic retrieval model dan DeBERTa v3 sebagai reranking model dalam sistem *chatbot* otomotif berbasis *Retrieval-augmented generation*. Selain itu, penelitian ini melakukan evaluasi menggunakan framework RAGAS untuk mengukur kualitas retrieval dan generation secara lebih komprehensif (Wang et al., 2024; Gao & When, 2023; James et al., 2024).

2. KAJIAN TEORITIS

Chatbot

Chatbot merupakan aplikasi berbasis kecerdasan buatan yang dirancang untuk melakukan interaksi dengan pengguna menggunakan bahasa alami melalui media teks maupun suara. Teknologi *chatbot* banyak digunakan sebagai sarana penyedia informasi karena mampu memberikan respons secara otomatis dan cepat. Perkembangan teknologi *Natural Language Processing* (NLP) memungkinkan *chatbot* memahami maksud pertanyaan pengguna serta menghasilkan jawaban yang lebih relevan sesuai konteks percakapan. Pada awal perkembangannya, *chatbot* umumnya dibangun menggunakan pendekatan berbasis aturan (rule-based) yang hanya mampu merespons pertanyaan sesuai pola yang telah ditentukan.

Seiring berkembangnya teknologi kecerdasan buatan, *chatbot* modern mulai memanfaatkan model pembelajaran mesin dan *Large Language Model* (LLM) sehingga mampu menghasilkan jawaban yang lebih fleksibel dan menyerupai komunikasi manusia. Dalam penelitian ini, *chatbot* digunakan sebagai media layanan informasi otomotif yang memungkinkan pengguna memperoleh informasi teknis kendaraan melalui percakapan berbasis bahasa alami (Elysia & Herianto, 2024; Sianipar et al., 2022; Lewis et al., 2020).

Large Language Model (LLM)

Large Language Model (LLM) merupakan model kecerdasan buatan yang dilatih menggunakan data teks dalam jumlah besar untuk memahami dan menghasilkan bahasa alami. Model ini mampu melakukan berbagai tugas pemrosesan bahasa, seperti menjawab pertanyaan, merangkum teks, menerjemahkan bahasa, serta menghasilkan konten berdasarkan konteks yang diberikan pengguna. Kemampuan LLM dalam memahami konteks percakapan menjadikannya komponen penting dalam pengembangan *chatbot* modern. Meskipun demikian, model bahasa besar masih memiliki kelemahan berupa hallucination, yaitu kondisi ketika model menghasilkan informasi yang terlihat benar tetapi tidak didukung oleh sumber informasi yang valid. Oleh karena itu, diperlukan pendekatan tambahan yang mampu menghubungkan model bahasa dengan sumber informasi yang terpercaya agar kualitas jawaban dapat ditingkatkan (Huang et al., 2025)

Retrieval-augmented generation (RAG)

Retrieval-augmented generation (RAG) merupakan pendekatan yang menggabungkan proses pencarian informasi (*retrieval*) dan pembangkitan jawaban (*Generation*). Pada metode ini, sistem terlebih dahulu mencari dokumen yang relevan terhadap pertanyaan pengguna, kemudian dokumen tersebut digunakan sebagai konteks tambahan dalam proses generasi jawaban oleh Large Language Model. Pendekatan RAG memiliki beberapa keunggulan dibandingkan model generatif murni. Pertama, jawaban yang dihasilkan lebih akurat karena didukung oleh dokumen referensi yang relevan. Kedua, risiko hallucination dapat dikurangi karena model bahasa tidak hanya bergantung pada pengetahuan hasil pelatihan. Ketiga, sistem dapat memanfaatkan informasi terbaru tanpa perlu melakukan pelatihan ulang model bahasa. Oleh karena itu, metode RAG banyak digunakan pada sistem tanya jawab, *chatbot*, dan layanan informasi berbasis kecerdasan buatan (Hoshi et al., 2023; Zhou et al., 2024).

Multilingual E5

Multilingual E5 merupakan model embedding yang dirancang untuk menghasilkan representasi semantik teks dalam berbagai bahasa. Model ini mengubah dokumen maupun pertanyaan pengguna menjadi vektor numerik yang merepresentasikan makna suatu teks

sehingga memungkinkan proses pencarian informasi dilakukan berdasarkan kesamaan makna, bukan hanya kesamaan kata. Keunggulan Multilingual E5 terletak pada kemampuannya memahami hubungan semantik antar kalimat dalam berbagai bahasa. Dengan kemampuan tersebut, sistem dapat menemukan dokumen yang relevan meskipun pengguna menggunakan kata atau kalimat yang berbeda dari dokumen yang tersimpan pada basis pengetahuan. Dalam penelitian ini, Multilingual E5 digunakan sebagai model embedding pada tahap *retrieval* untuk meningkatkan kualitas pencarian dokumen (Wang et al., 2024).

DeBERTa v3

DeBERTa v3 (*Decoding-enhanced Bidirectional Encoder Representation from Transformers Version 3*) merupakan model transformer yang dikembangkan untuk meningkatkan kemampuan pemahaman konteks bahasa. Model ini memiliki performa yang baik dalam berbagai tugas pemrosesan bahasa alami, termasuk klasifikasi teks, pencarian informasi, dan penilaian relevansi dokumen. Pada penelitian ini, DeBERTa v3 digunakan sebagai model reranking. Setelah proses *retrieval* menghasilkan sejumlah dokumen kandidat, DeBERTa v3 melakukan evaluasi ulang terhadap hubungan antara pertanyaan pengguna dan dokumen yang ditemukan. Hasil evaluasi tersebut digunakan untuk menentukan urutan dokumen yang paling relevan sehingga konteks yang diberikan kepada Large Language Model menjadi lebih berkualitas (He et al., 2023).

RAGAS

RAGAS (*Retrieval-augmented generation Assessment*) merupakan framework evaluasi yang dirancang khusus untuk mengukur performa sistem berbasis *Retrieval-augmented generation*. Framework ini memungkinkan evaluasi dilakukan secara otomatis terhadap kualitas *retrieval* maupun kualitas jawaban yang dihasilkan oleh sistem. Beberapa metrik yang umum digunakan dalam RAGAS meliputi *Faithfulness*, *Answer Relevancy*, *Context Precision*, dan *Context Recall*. *Faithfulness* digunakan untuk mengukur kesesuaian jawaban terhadap dokumen konteks, *Answer Relevancy* digunakan untuk mengukur relevansi jawaban terhadap pertanyaan pengguna, sedangkan *Context Precision* dan *Context Recall* digunakan untuk mengevaluasi kualitas dokumen yang diperoleh pada tahap *retrieval*. Dalam penelitian ini, RAGAS digunakan sebagai metode utama untuk mengevaluasi performa *chatbot* otomatis yang dikembangkan (Es et al., 2024).

3. METODE PENELITIAN

Penelitian ini merupakan penelitian terapan (*applied research*) yang berfokus pada pengembangan sistem *chatbot* otomotif berbasis *Retrieval-augmented generation* (RAG). Penelitian dilakukan untuk menghasilkan sistem yang mampu memberikan jawaban yang relevan, akurat, dan berbasis referensi terhadap berbagai permasalahan otomotif yang diajukan oleh pengguna. Pendekatan yang digunakan memanfaatkan kombinasi *semantic retrieval* dan Large Language Model untuk meningkatkan kualitas jawaban yang dihasilkan.

Metode Pengembangan Sistem

Pengembangan sistem dilakukan dengan mengintegrasikan beberapa komponen utama yang mendukung proses *Retrieval-augmented generation*. Komponen tersebut meliputi Multilingual E5 sebagai model embedding (Wang et al., 2024), FAISS sebagai vector database (Pan et al., 2024), DeBERTa v3 sebagai model reranking (He et al., 2023), serta Qwen2.5 sebagai *Large Language Model* (Huei et al., 2024), yang bertugas menghasilkan jawaban akhir. Tahapan pengembangan sistem dimulai dari pengumpulan dokumen otomotif, preprocessing data, proses embedding, penyimpanan vektor pada vector database, *retrieval* dokumen, reranking dokumen, hingga proses generasi jawaban. Seluruh tahapan dirancang untuk memastikan bahwa jawaban yang dihasilkan didukung oleh dokumen referensi yang relevan.

Teknik Pengumpulan Data

Pengumpulan data dilakukan untuk memperoleh basis pengetahuan yang digunakan oleh *chatbot* dalam menjawab pertanyaan pengguna. Teknik pengumpulan data yang digunakan meliputi: Studi Literatur dilakukan dengan mengumpulkan berbagai referensi yang berkaitan dengan sistem otomotif, perawatan kendaraan, komponen kendaraan, diagnosa kerusakan kendaraan, dan teknologi *Retrieval-augmented generation*. Referensi diperoleh dari buku, jurnal ilmiah, dokumentasi teknis, dan sumber informasi terpercaya lainnya; Studi Dokumentasi dilakukan dengan mengumpulkan dokumen otomotif yang akan digunakan sebagai basis pengetahuan sistem. Dokumen tersebut kemudian diproses dan disesuaikan agar dapat digunakan dalam proses *retrieval* dan *Generation*.

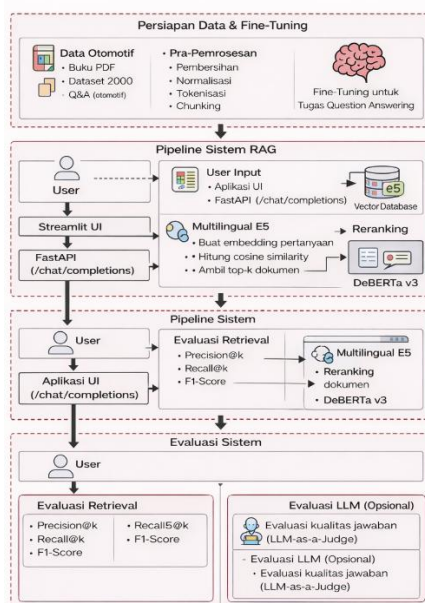
Perancangan dan Implementasi Sistem

Sistem *chatbot* dikembangkan menggunakan pendekatan *Retrieval-augmented generation* yang menggabungkan proses pencarian dokumen dan generasi jawaban. Arsitektur sistem terdiri atas beberapa komponen utama yang saling terintegrasi untuk menghasilkan jawaban yang lebih akurat dan kontekstual. Tahap awal dimulai dengan pengumpulan dokumen otomotif yang digunakan sebagai basis pengetahuan. Dokumen yang telah diperoleh kemudian melalui proses preprocessing untuk membersihkan data dan menyeragamkan format teks.

Setelah itu, dokumen diproses menggunakan teknik chunking untuk membagi dokumen menjadi beberapa bagian yang lebih kecil sehingga proses *retrieval* dapat dilakukan secara lebih efektif.

Implementasi *Chatbot* Berbasis RAG

Implementasi *chatbot* pada penelitian ini menggunakan pendekatan *Retrieval-augmented generation* yang mengintegrasikan *semantic retrieval* dan Large Language Model. Arsitektur sistem yang digunakan ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Sistem Chatbot Berbasis Retrieval-augmented generation

Tahapan implementasi sistem dijelaskan sebagai berikut:

Preprocessing Data

Tahap preprocessing dilakukan untuk membersihkan dokumen dari karakter yang tidak diperlukan, memperbaiki struktur teks, serta menyeragamkan format data. Tahap ini bertujuan meningkatkan kualitas dokumen sebelum digunakan pada proses *retrieval*.

Text Chunking

Dokumen yang telah diproses kemudian dibagi menjadi beberapa bagian menggunakan teknik chunking. Proses ini dilakukan agar sistem dapat menemukan informasi yang lebih spesifik dan relevan terhadap pertanyaan pengguna.

Vector Embedding

Setiap potongan dokumen diubah menjadi representasi vektor menggunakan model Multilingual E5. Representasi vektor tersebut memungkinkan sistem memahami hubungan semantik antar dokumen dan pertanyaan pengguna.

Penyimpanan Vector Database

Embedding yang dihasilkan disimpan ke dalam FAISS sebagai vector database. Penyimpanan ini memungkinkan proses pencarian dokumen dilakukan secara efisien berdasarkan tingkat kemiripan vektor. Penyimpanan embedding pada FAISS memungkinkan proses pencarian dokumen dilakukan secara efisien melalui mekanisme vector similarity search [10].

Input Pertanyaan Pengguna

Pengguna mengajukan pertanyaan melalui antarmuka *chatbot* berbasis web. Pertanyaan yang diterima kemudian diproses menggunakan model embedding yang sama untuk menghasilkan representasi vektor pertanyaan.

Semantic retrieval

Sistem melakukan pencarian dokumen yang paling relevan menggunakan metode cosine similarity antara embedding pertanyaan dan embedding dokumen yang tersimpan pada vector database. Hasil *retrieval* berupa sejumlah dokumen kandidat yang memiliki tingkat kemiripan tertinggi.

Re-ranking Dokumen

Dokumen hasil *retrieval* diproses kembali menggunakan DeBERTa v3 untuk menentukan urutan relevansi yang lebih akurat. Tahap ini bertujuan meningkatkan kualitas konteks yang akan digunakan pada proses generasi jawaban.

Response Generation

Dokumen terbaik hasil reranking digunakan sebagai konteks tambahan oleh model Qwen2.5 untuk menghasilkan jawaban yang sesuai dengan pertanyaan pengguna. Jawaban yang dihasilkan kemudian dikirimkan kembali melalui antarmuka *chatbot*.

Teknik Pengujian Sistem

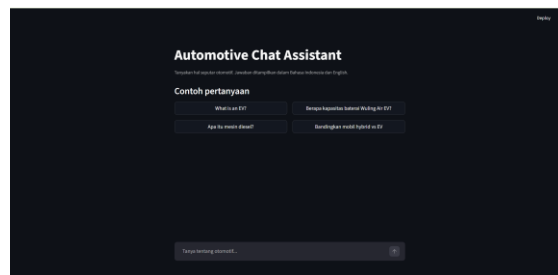
Pengujian dilakukan untuk mengukur performa *retrieval* dan kualitas jawaban yang dihasilkan sistem *chatbot*. Evaluasi *retrieval* dilakukan menggunakan metrik Precision, Recall, dan F1-Score untuk mengetahui kemampuan sistem dalam menemukan dokumen yang relevan (Moffat, 2022). Selain itu, evaluasi *Generation* dilakukan menggunakan framework RAGAS yang terdiri atas metrik *Faithfulness*, *Answer Relevancy*, *Context Precision*, dan *Context Recall*. Evaluasi ini bertujuan mengukur tingkat kesesuaian jawaban terhadap dokumen konteks serta relevansi jawaban terhadap pertanyaan pengguna. Hasil pengujian digunakan sebagai dasar untuk menilai efektivitas implementasi *Retrieval-augmented generation* pada sistem *chatbot* otomotif yang dikembangkan.

4. HASIL DAN PEMBAHASAN

Hasil Implementasi Sistem Chatbot Berbasis RAG

Penelitian ini menghasilkan sistem *chatbot* otomotif berbasis web yang menerapkan pendekatan *Retrieval-augmented generation* (RAG) untuk memberikan jawaban yang relevan, akurat, dan berbasis referensi. Sistem dikembangkan dengan mengintegrasikan Multilingual E5 sebagai model embedding, FAISS sebagai vector database, DeBERTa v3 sebagai model reranking, serta Qwen2.5 sebagai Large Language Model yang bertugas menghasilkan jawaban akhir.

Implementasi sistem dilakukan dengan menghubungkan seluruh komponen ke dalam satu alur kerja yang memungkinkan proses pencarian dokumen dan generasi jawaban berlangsung secara otomatis. Pengguna dapat mengajukan pertanyaan mengenai berbagai permasalahan otomotif menggunakan bahasa alami melalui antarmuka *chatbot* berbasis web. Sistem kemudian melakukan pencarian dokumen yang relevan sebelum menghasilkan jawaban berdasarkan konteks yang ditemukan.



Gambar 2. Antarmuka *Chatbot* Otomotif

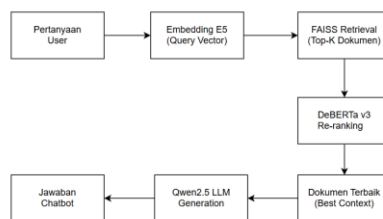
Pada tahap *retrieval*, pertanyaan pengguna diubah menjadi representasi vektor menggunakan Multilingual E5. Selanjutnya sistem melakukan pencarian dokumen pada vector database menggunakan metode cosine similarity. Dokumen yang berhasil ditemukan kemudian diproses kembali menggunakan DeBERTa v3 untuk menentukan urutan relevansi yang lebih akurat. Dokumen dengan tingkat relevansi tertinggi digunakan sebagai konteks tambahan dalam proses generasi jawaban oleh model Qwen2.5.

Berdasarkan hasil implementasi, sistem mampu menjalankan seluruh tahapan *Retrieval-augmented generation* dengan baik, mulai dari proses *retrieval*, reranking, hingga *Generation*. Integrasi tersebut memungkinkan *chatbot* menghasilkan jawaban yang tidak hanya berdasarkan pengetahuan bawaan model bahasa, tetapi juga memanfaatkan dokumen referensi yang tersedia pada basis pengetahuan otomotif.

Hasil Implementasi *Retrieval* dan *Re-ranking*

Salah satu komponen utama dalam penelitian ini adalah penerapan *semantic retrieval* menggunakan Multilingual E5 dan reranking menggunakan DeBERTa v3. Pendekatan ini digunakan untuk meningkatkan kualitas dokumen yang digunakan sebagai konteks dalam proses generasi jawaban.

Pada tahap *retrieval*, sistem mencari dokumen berdasarkan kesamaan makna antara pertanyaan pengguna dan dokumen yang tersimpan dalam basis pengetahuan. Pendekatan ini memungkinkan sistem menemukan dokumen yang relevan meskipun pengguna menggunakan kata atau kalimat yang berbeda dari dokumen sumber. Selanjutnya, dokumen hasil *retrieval* diproses menggunakan DeBERTa v3 untuk menentukan urutan relevansi yang lebih akurat. Tahap reranking dilakukan karena dokumen yang memiliki nilai similarity tinggi belum tentu merupakan dokumen yang paling sesuai dengan kebutuhan pengguna. Dengan adanya proses reranking, sistem dapat memilih dokumen terbaik yang akan digunakan sebagai konteks pada tahap *Generation*.



Gambar 3. Implementasi *Retrieval* dan *Re-ranking* Menggunakan *Multilingual E5* dan *DeBERTa v3*

Hasil implementasi menunjukkan bahwa kombinasi *semantic retrieval* dan reranking mampu meningkatkan kualitas konteks yang digunakan oleh Large Language Model. Dengan konteks yang lebih relevan, jawaban yang dihasilkan menjadi lebih akurat dan sesuai dengan kebutuhan informasi pengguna.

Hasil Evaluasi *Retrieval*

Evaluasi *retrieval* dilakukan untuk mengukur kemampuan sistem dalam menemukan dokumen yang relevan terhadap pertanyaan pengguna. Pengujian dilakukan menggunakan metrik *Precision*, *Recall*, dan *F1-Score*.

Tabel 1. Hasil *Evaluasi Retrieval*

No	Metrik	Nilai
1	Precision	0,192
2	Recall	0,705
3	F1-Score	0,290
4	<i>Retrieval Score</i>	0,714

Nilai Recall yang lebih tinggi dibandingkan Precision menunjukkan bahwa sistem cenderung mampu menemukan sebagian besar dokumen yang relevan meskipun masih terdapat beberapa dokumen yang kurang relevan pada hasil *retrieval* awal. Kondisi tersebut merupakan karakteristik umum pada sistem *semantic retrieval* yang mengutamakan cakupan informasi sebelum dilakukan proses reranking menggunakan DeBERTa v3.

Berdasarkan hasil pengujian, sistem memperoleh nilai Recall yang relatif tinggi. Hasil tersebut menunjukkan bahwa sebagian besar dokumen yang relevan berhasil ditemukan oleh mekanisme *retrieval* yang digunakan. Nilai *Retrieval Score* yang diperoleh juga menunjukkan bahwa proses pencarian dokumen mampu menyediakan konteks yang cukup baik untuk mendukung proses generasi jawaban.

Meskipun nilai Precision masih lebih rendah dibandingkan Recall, penggunaan Multilingual E5 terbukti mampu meningkatkan kemampuan sistem dalam menemukan dokumen berdasarkan kesamaan makna. Pendekatan ini lebih efektif dibandingkan pencarian berbasis kata kunci karena mampu memahami hubungan semantik antara pertanyaan dan dokumen.

Hasil Evaluasi *Chatbot* Menggunakan RAGAS

Evaluasi kualitas jawaban dilakukan menggunakan framework RAGAS untuk mengukur performa sistem *Retrieval-augmented generation*. Pengujian difokuskan pada metrik Faithfulness, Answer Relevancy, dan Context Precision.

Tabel 2. Hasil *Evaluasi RAGAS*

No	Metrik	Nilai
1	Faithfulness	1,000
2	Answer Relevancy	0,883
3	Context Precision	1,000

Berdasarkan hasil evaluasi, sistem memperoleh nilai Faithfulness sebesar 1,000. Nilai tersebut menunjukkan bahwa jawaban yang dihasilkan *chatbot* sepenuhnya didukung oleh dokumen konteks yang digunakan selama proses generasi jawaban. Dengan kata lain, sistem mampu menghasilkan jawaban yang konsisten dengan informasi yang terdapat pada basis pengetahuan.

Nilai *Answer Relevancy* sebesar 0,883 menunjukkan bahwa jawaban yang dihasilkan memiliki tingkat kesesuaian yang tinggi terhadap pertanyaan pengguna. Hasil ini mengindikasikan bahwa kombinasi *retrieval* dan reranking mampu membantu sistem memahami kebutuhan informasi pengguna secara lebih baik. Selain itu, nilai Context Precision sebesar 1,000 menunjukkan bahwa dokumen yang digunakan sebagai konteks merupakan

dokumen yang relevan terhadap pertanyaan yang diajukan. Hasil tersebut membuktikan bahwa penggunaan DeBERTa v3 pada tahap reranking berhasil meningkatkan kualitas dokumen yang digunakan dalam proses generasi jawaban.

Analisis Hasil Jawaban *Chatbot*

Untuk mengetahui kualitas jawaban yang dihasilkan sistem secara lebih mendalam, dilakukan pengujian menggunakan beberapa pertanyaan yang mewakili kasus umum pada domain otomotif. Hasil pengujian menunjukkan bahwa sistem mampu memberikan jawaban yang relevan ketika informasi yang dibutuhkan tersedia pada basis pengetahuan.

Tabel 3. Ringkasan Analisis Hasil Jawaban *Chatbot*

No	Pertanyaan	Hasil
1	Apa penyebab mesin mobil sulit dihidupkan?	Berhasil
2	Apa fungsi radiator pada kendaraan?	Berhasil
3	Kenapa mobil saya bermasalah?	Kurang Optimal

Berdasarkan Tabel 3, sistem menunjukkan performa yang baik pada pertanyaan yang memiliki konteks dan informasi yang jelas, sedangkan performa sistem cenderung menurun ketika pertanyaan yang diberikan bersifat umum atau ambigu sehingga konteks yang diperoleh menjadi kurang spesifik.

Salah satu contoh pertanyaan yang berhasil dijawab dengan baik adalah:

Pertanyaan:

"Apa penyebab mesin mobil sulit dihidupkan?"

Sistem mampu menjelaskan beberapa kemungkinan penyebab, seperti kondisi aki yang lemah, kerusakan pada sistem starter, gangguan pada sistem bahan bakar, serta masalah pada sistem pengapian. Jawaban yang diberikan sesuai dengan informasi yang terdapat pada dokumen referensi sehingga memperoleh nilai faithfulness yang tinggi.

Contoh lain adalah pertanyaan:

Pertanyaan:

"Apa fungsi radiator pada kendaraan?"

Chatbot mampu memberikan penjelasan yang lengkap mengenai fungsi radiator sebagai komponen sistem pendingin yang bertugas menjaga suhu kerja mesin agar tetap stabil. Jawaban yang dihasilkan juga didukung oleh konteks yang relevan dari basis pengetahuan.

Meskipun demikian, masih terdapat beberapa kondisi di mana jawaban yang dihasilkan belum sepenuhnya optimal. Sebagai contoh, ketika pengguna memberikan pertanyaan yang terlalu umum atau ambigu, sistem terkadang menghasilkan jawaban yang kurang spesifik.

Pertanyaan:

"Kenapa mobil saya bermasalah?"

Pada kasus tersebut, sistem kesulitan menentukan konteks kerusakan yang dimaksud karena pertanyaan tidak mengandung informasi yang cukup. Kondisi ini menyebabkan jawaban yang dihasilkan masih bersifat umum dan belum mampu mengidentifikasi jenis kerusakan kendaraan secara spesifik. Hal tersebut menunjukkan bahwa kualitas jawaban sangat dipengaruhi oleh tingkat kejelasan dan kelengkapan informasi yang diberikan pengguna.

Selain itu, kualitas jawaban juga sangat dipengaruhi oleh kelengkapan dokumen pada basis pengetahuan. Jika informasi yang dibutuhkan tidak tersedia pada dokumen referensi, sistem cenderung memberikan jawaban yang lebih singkat atau kurang detail dibandingkan pertanyaan yang memiliki dukungan dokumen yang memadai.

Keunggulan dan Kelemahan Sistem

Berdasarkan hasil implementasi dan evaluasi yang telah dilakukan, sistem chatbot otomotif yang dikembangkan memiliki beberapa keunggulan. Sistem mampu melakukan pencarian informasi berdasarkan kesamaan makna menggunakan Multilingual E5 sehingga tidak bergantung pada kesamaan kata secara langsung. Penggunaan DeBERTa v3 pada tahap *reranking* mampu meningkatkan kualitas dokumen konteks yang digunakan oleh *Large Language Model*. Sistem memperoleh nilai *Faithfulness* dan *Context Precision* yang tinggi sehingga jawaban yang dihasilkan didukung oleh dokumen referensi yang relevan. Implementasi *retrieval-augmented generation* membantu mengurangi potensi *hallucination* yang umum terjadi pada *Large Language Model*. Selain memiliki beberapa keunggulan, sistem juga masih memiliki beberapa keterbatasan. Kualitas jawaban sangat bergantung pada kelengkapan dan kualitas dokumen yang tersedia pada basis pengetahuan. Pertanyaan yang terlalu umum atau ambigu masih dapat menghasilkan jawaban yang kurang spesifik. Nilai *Precision* pada proses *retrieval* masih relatif rendah sehingga masih terdapat kemungkinan dokumen yang kurang relevan muncul pada hasil pencarian awal. Sistem saat ini hanya memanfaatkan data berbasis teks dan belum mendukung data multimodal seperti gambar atau video kerusakan kendaraan. Meskipun demikian, keterbatasan tersebut dapat menjadi peluang pengembangan pada penelitian selanjutnya melalui integrasi data multimodal dan perluasan basis pengetahuan sehingga sistem mampu memberikan layanan informasi otomotif yang lebih komprehensif.

Pembahasan

Hasil penelitian menunjukkan bahwa penerapan *Retrieval-augmented generation* mampu meningkatkan kualitas layanan informasi otomotif melalui integrasi proses *retrieval* dan *Generation*. Penggunaan Multilingual E5 memungkinkan sistem melakukan pencarian dokumen berdasarkan kesamaan makna sehingga informasi yang ditemukan tidak bergantung

pada kesamaan kata secara langsung (Wang et al., 2024). Implementasi DeBERTa v3 pada tahap reranking memberikan kontribusi penting dalam meningkatkan kualitas konteks yang digunakan oleh Large Language Model (He et al., 2023). Dengan memilih dokumen yang lebih relevan, sistem mampu menghasilkan jawaban yang lebih akurat dan sesuai dengan kebutuhan pengguna.

Hasil evaluasi menggunakan framework RAGAS menunjukkan bahwa sistem memiliki tingkat *Faithfulness* dan *Context Precision* yang sangat tinggi. Hal tersebut menunjukkan bahwa jawaban yang dihasilkan tidak hanya relevan tetapi juga didukung oleh dokumen referensi yang sesuai. Dengan demikian, pendekatan yang digunakan mampu membantu mengurangi potensi hallucination yang sering ditemukan pada Large Language Model (Huang et al., 2025; Es et al., 2024). Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi Multilingual E5, FAISS, DeBERTa v3, dan Qwen2.5 dapat digunakan secara efektif untuk membangun *chatbot* otomotif yang mampu memberikan jawaban yang relevan, akurat, dan berbasis referensi. Hasil tersebut menunjukkan bahwa pendekatan *Retrieval-augmented generation* memiliki potensi besar untuk diterapkan pada layanan informasi otomotif berbasis kecerdasan buatan. Hasil penelitian ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa pendekatan Retrieval-Augmented Generation mampu meningkatkan kualitas jawaban chatbot pada berbagai domain, seperti layanan pendidikan, kesehatan, dan sistem informasi klinik (Elysia & Herianto, 2024; Zhou et al., 2024; Samudra et al., 2025; Surya et al., 2026). Namun demikian, penelitian ini menunjukkan bahwa kombinasi Multilingual E5 dan DeBERTa v3 memberikan performa yang baik pada domain otomotif dengan menghasilkan nilai *Faithfulness* dan *Context Precision* yang tinggi. Temuan tersebut menunjukkan bahwa pendekatan *semantic retrieval* dan reranking dapat menjadi solusi efektif untuk meningkatkan kualitas layanan informasi otomotif berbasis kecerdasan buatan.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan *chatbot* otomotif berbasis *Retrieval-augmented generation* dengan mengintegrasikan Multilingual E5 sebagai semantic embedding dan DeBERTa v3 sebagai reranking model. Kombinasi kedua metode tersebut memungkinkan sistem memperoleh dokumen yang lebih relevan sehingga mampu menghasilkan jawaban yang lebih kontekstual dan berbasis referensi. Temuan penelitian menunjukkan bahwa penerapan *semantic retrieval* dan reranking mampu meningkatkan kualitas jawaban *chatbot* serta membantu mengurangi potensi hallucination pada Large Language Model. Dengan demikian, tujuan penelitian untuk menghasilkan sistem *chatbot* otomotif yang mampu memberikan

informasi yang lebih akurat, relevan, dan kontekstual berhasil dicapai. Penelitian selanjutnya dapat difokuskan pada pengembangan basis pengetahuan yang lebih luas, integrasi data multimodal, serta optimasi performa sistem untuk mendukung implementasi pada lingkungan produksi dengan jumlah pengguna yang lebih besar.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Tarumanagara, khususnya Program Studi Teknik Informatika Fakultas Teknologi Informasi, yang telah memberikan dukungan akademik dan fasilitas selama pelaksanaan penelitian ini.

Penulis juga menyampaikan apresiasi kepada dosen pembimbing yang telah memberikan arahan, masukan, serta bimbingan selama proses penelitian dan penyusunan artikel ini. Ucapan terima kasih turut disampaikan kepada seluruh narasumber dan pihak terkait yang telah memberikan informasi, masukan, dan dukungan sehingga penelitian mengenai pengembangan *chatbot* otomotif berbasis *Retrieval-augmented generation* dapat diselesaikan dengan baik. Semoga hasil penelitian ini dapat memberikan kontribusi bagi pengembangan teknologi *chatbot* dan layanan informasi otomotif berbasis kecerdasan buatan.

DAFTAR REFERENSI

- Moffat, A. (2022). *Batch evaluation metrics in information retrieval: Measures, scales, and meaning*. *IEEE Access*, 10. <https://doi.org/10.1109/ACCESS.2022.3211668>
- Hui, B., et al. (2024). *Qwen2.5-Coder technical report*. *arXiv*. <http://arxiv.org/abs/2409.12186>
- Samudra, G., Turmudi Zy, A., & Ermanto. (2025). *Implementation of retrieval augmented generation (RAG) in the design of digestive health chatbot*. *JSAI: Journal Scientific and Applied Informatics*, 8(1).
- Pan, J. J., Wang, J., & Li, G. (2024). *Survey of vector database management systems*. *The VLDB Journal*, 33(5). <https://doi.org/10.1007/s00778-024-00864-x>
- Huang, L., et al. (2025). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. *ACM Transactions on Information Systems*, 43(2). <https://doi.org/10.1145/3703155>
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 text embeddings: A technical report*. *arXiv*. <http://arxiv.org/abs/2402.05672>
- He, P., Gao, J., & Chen, W. (2023). *DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing*. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.
- Surya, P., Permana, J., Luh, N., & Candrawengi, P. I. (2026). *Implementasi chatbot berbasis RAG pada sistem informasi rawat jalan klinik*. *Jurnal Publikasi Teknik Informatika*, 5, 52–69. <https://doi.org/10.55606/jupti.v5i2.7063>

- Zhou, Q., et al. (2024). *GastroBot: A Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation*. *Frontiers in Medicine*, 11. <https://doi.org/10.3389/fmed.2024.1392555>
- Elysia, S., & Herianto. (2024). *Chatbot berbasis retrieval augmented generation (RAG) untuk peningkatan layanan informasi sekolah*. *Journal TIFDA (Technology Information and Data Analytic)*, 1(2). <https://doi.org/10.70491/tifda.v1i2.52>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). *RAGAS: Automated evaluation of retrieval augmented generation*. In *Proceedings of the EACL 2024: 18th Conference of the European Chapter of the Association for Computational Linguistics, System Demonstrations*. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Sianipar, Y. P., Mawardi, V. C., & Sutrisno, T. (2022). *Penggunaan Apriori pada rekomendasi paket menu dan dilengkapi fitur chatbot*. *Jurnal Ilmu Komputer dan Sistem Informasi*, 10(1). <https://doi.org/10.24912/jiksi.v10i1.17821>
- Hoshi, Y., et al. (2023). *Retrieval-augmented generation for large language models: A survey*.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). *Dense passage retrieval for open-domain question answering*. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- 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.