

Penerapan Metode *Retrieval-Augmented Generation (RAG)* Pada *Chatbot E-Commerce* Berbasis Gemini Ai

Tarisa Ramadhani¹, Noora Qotrun Nada², Nugroho Dwi S³
Fakultas Teknik Dan Informatika, Program Informatika
Universitas Persatuan Guru Republik Indonesia Semarang
sasarisa079@gmail.com

Abstrak

Perkembangan *e-commerce* yang pesat telah meningkatkan harapan untuk sistem layanan pelanggan yang cerdas dan sadar konteks. *Chatbot* berbasis aturan tradisional terbukti tidak memadai dalam memenuhi permintaan yang semakin meningkat untuk respons yang akurat, cepat, dan relevan secara kontekstual. Penelitian ini mengusulkan pengembangan sistem *chatbot* berbasis Gemini AI yang terintegrasi dengan pendekatan *Retrieval-Augmented Generation (RAG)*. Sistem ini mengambil dokumen-dokumen relevan dari basis data internal (katalog produk, kebijakan, riwayat pelanggan) dan menggunakan model LLaMA sebagai generator untuk menghasilkan jawaban yang faktual dan sejalan secara semantis. Dataset yang digunakan mencakup 214 pasangan percakapan yang telah dibersihkan yang bersumber dari Kaggle, diproses menggunakan *Sentence-BERT* untuk *embedding* kalimat. Sistem *chatbot* dievaluasi melalui metrik pengambilan dan generatif. Hasil menunjukkan kinerja tinggi dengan *Mean Reciprocal Rank (MRR)* sebesar 0,83, *Exact Match (EM)* sebesar 100%, *F1 Score* sebesar 82,05%, dan *Similarity Semantis* sebesar 97,45%, Tingkat kesetiaan 91,67% dan relevansi jawaban 94,21%. Temuan ini menunjukkan bahwa integrasi Gemini AI dengan RAG dapat secara signifikan meningkatkan akurasi, fakta, dan relevansi konteks dari *respons chatbot* di domain dinamis yang didorong oleh data seperti *e-commerce*.

Kata kunci: *chatbot, e-commerce, Gemini AI, Retrieval-Augmented Generation, kesamaan semantic.*

Abstract

The rapid growth of e-commerce has raised expectations for intelligent and context-aware customer service systems. Traditional rule-based chatbots have proven inadequate in meeting the growing demand for accurate, fast, and contextually relevant responses. This research proposes the development of a Gemini AI-based chatbot system integrated with a Retrieval-Augmented Generation (RAG) approach. The system retrieves relevant documents from internal databases (product catalogues, policies, customer histories) and uses the LLaMA model as a generator to produce factual and semantically aligned answers. The dataset used includes 214 cleaned conversation pairs sourced from Kaggle, processed using Sentence-BERT for sentence embedding. The chatbot system is evaluated through retrieval and generative metrics. The results show high performance with a Mean Reciprocal Rank (MRR) of 0.83, an Exact Match (EM) of 100%, an F1 Score of 82.05%, a Semantic Similarity of 97.45%, a Fidelity Rate of 91.67%, and an answer relevance of 94.21%. These findings indicate that the integration of Gemini AI with RAG can significantly improve the accuracy, factuality, and contextual relevance of chatbot responses in dynamic, data-driven domains such as e-commerce.

Keywords: *chatbots, e-commerce, Gemini AI, Retrieval-Augmented Generation, semantic similarity.*

PENDAHULUAN

Dalam beberapa dekade terakhir, transformasi digital telah mengubah secara mendasar cara konsumen berinteraksi dengan bisnis. *E-commerce* sebagai salah satu sektor yang paling terdampak mengalami pertumbuhan pesat, terutama seiring meningkatnya penggunaan internet dan perangkat pintar. Menurut data Kementerian Komunikasi dan Informatika, sektor *e-commerce* berkontribusi sebesar 64,5% terhadap total ekonomi digital Indonesia pada tahun 2022 [1]. Fenomena ini mendorong peningkatan ekspektasi konsumen terhadap layanan pelanggan yang tidak hanya cepat, tetapi juga akurat, personal, dan kontekstual.

Chatbot adalah salah satu solusi digital yang berkembang pesat dalam menjawab kebutuhan layanan pelanggan secara otomatis. *Chatbot* awalnya dikembangkan dengan pendekatan berbasis aturan (*rulebased*), namun kini telah bertransformasi menjadi sistem berbasis *Natural Language Processing* (NLP) dan *Large Language Model* (LLM) seperti GPT, BERT, dan Gemini AI [2]. Model generatif ini mampu menghasilkan bahasa alami yang menyerupai manusia serta memberikan pengalaman percakapan yang lebih interaktif dan [3]. Dalam konteks bisnis, *chatbot* dinilai mampu meningkatkan efisiensi operasional karena dapat memberikan respons instan dan aktif selama 24 jam.

Meskipun demikian, penggunaan LLM dalam pengembangan *chatbot* tidak lepas dari berbagai tantangan. Salah satu kendala utama adalah fenomena *hallucination*, yaitu ketika model menghasilkan jawaban yang tidak akurat atau tidak berdasar fakta [4]. Selain itu, model generatif cenderung memiliki

keterbatasan dalam memori kontekstual jangka Panjang dan tidak terhubung secara langsung ke sumber data terkini. Hal ini menjadi masalah serius dalam domain seperti *e-commerce* yang sangat dinamis dan data-driven. Integrasi metode RAG dengan Gemini AI, model generatif multimodal dari *Google DeepMind*, menghadirkan peluang baru dalam pengembangan *chatbot* yang cerdas, adaptif, dan informatif. Gemini AI memiliki keunggulan dalam menangani data multimodal seperti *teks* dan gambar secara terpadu, sehingga sangat cocok diterapkan dalam lingkungan *e-commerce* yang sarat dengan konten visual dan deskriptif [5]. Dengan menggabungkan kemampuan retrieval dan generasi dalam satu sistem, *chatbot* dapat memberikan jawaban yang tidak hanya natural tetapi juga berdasarkan bukti faktual dari data internal perusahaan seperti katalog produk, kebijakan toko, serta riwayat percakapan pelanggan. generator berbasis Gemini AI, dengan basis data yang disesuaikan terhadap kebutuhan informasi dalam layanan *e-commerce*. Sistem kemudian diuji menggunakan dataset katalog produk, dokumen kebijakan, dan riwayat interaksi pelanggan.

Berdasarkan uraian tersebut, maka rumusan masalah dalam penelitian ini adalah bagaimana metode RAG dapat meningkatkan akurasi dan relevansi jawaban *chatbot* generatif berbasis Gemini AI dalam domain *e-commerce*. Rencana pemecahan masalah dilakukan dengan membangun arsitektur *chatbot* yang menggabungkan modul *retriever*. Mengenai teknik RAG, kami mengumpulkan rincian tentang LLM dasar, sumber pencarian, dan strategi yang digunakan yang digunakan pada tahap pra-

pengambilan, pengambilan, dan pasca-pengambilan [6].

Rumusan Masalah

Bagaimana penerapan metode Retrieval-Augmented Generation (RAG) dapat meningkatkan akurasi dan relevansi jawaban chatbot generatif berbasis Gemini AI dalam domain e-commerce?

Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengembangkan *chatbot e-commerce* berbasis Gemini AI dengan pendekatan RAG, menganalisis performa *chatbot* dalam memberikan jawaban yang faktual dan kontekstual, Menunjukkan keunggulan metode RAG dibandingkan pendekatan *chatbot* konvensional berbasis LLM murni.

Manfaat Penelitian

1. Meningkatkan Kualitas Layanan Pelanggan
Chatbot yang dikembangkan mampu memberikan jawaban yang lebih faktual, relevan, dan sesuai konteks, sehingga meningkatkan pengalaman pengguna dalam e-commerce.
2. Mengurangi Hallucination pada Model Generatif
Integrasi metode RAG membantu mengurangi jawaban yang tidak akurat karena chatbot mengacu pada dokumen internal sebagai sumber data.
3. Sebagai Referensi Pengembangan Chatbot Cerdas
Penelitian ini menjadi acuan bagi pengembangan chatbot berbasis AI di sektor lain yang memerlukan interaksi kontekstual dan berbasis data real-time.

Tinjauan Pustaka

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan layanan pelanggan cerdas di sektor *e-commerce* serta menjadi referensi dalam penerapan LLM generatif yang terhubung dengan data dinamis secara *real-time*. Untuk mengatasi keterbatasan tersebut, metode *Retrieval-Augmented Generation* (RAG) diperkenalkan sebagai pendekatan hibrida yang menggabungkan kemampuan pencarian dokumen eksternal (*retriever*) dengan proses generatif (*generator*) [7]. RAG memungkinkan sistem mengakses informasi faktual dari basis data atau dokumen yang relevan, yang kemudian digunakan sebagai konteks saat menghasilkan jawaban [8]. Pendekatan ini dinilai mampu mengurangi tingkat *hallucination*, meningkatkan relevansi jawaban, serta mengakomodasi kebutuhan informasi yang terus berubah.

Studi ini menemukan bahwa Persepsi kemudahan penggunaan, Persepsi kegunaan, Kompatibilitas, Sikap terhadap Iklan, dan Pengaruh Sosial Pengaruh berkontribusi pada Sikap Generasi *Baby boomer* terhadap *chatbot* AI. Secara deskriptif, mayoritas responden *baby boomer* menghabiskan 2-4 jam di di media sosial dan aktivitas yang paling umum saat menggunakan media sosial adalah tentang hiburan [9].

LANDASAN TEORI

CHATBOT

Chatbot adalah perangkat lunak komputer yang dibuat untuk meniru teks dan intelektual berbasis audio percakapan dengan satu atau lebih manusia. *Chatbot* juga dapat diartikan sebagai komputer sistem yang memungkinkan manusia untuk berinteraksi dengan dengan komputer

menggunakan bahasa alami manusia. *Chatbot* berasal dari dua kata, yaitu “*chat*” dan “*bot*”. *Chat* dapat diibaratkan sebagai sebuah media tertulis aktivitas komunikasi dalam dunia komputer. *Bot* adalah sebuah program komputer dengan beberapa data poin yang, ketika diberi masukan, akan menghasilkan sebuah keluaran sebagai tanggapan [10].

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) adalah sebuah pendekatan arsitektur yang memadukan mekanisme pencarian informasi dengan kemampuan generatif dari model bahasa. Konsep ini diperkenalkan oleh Lewis et al [11]. sebagai strategi untuk meningkatkan kualitas respons yang dihasilkan oleh sistem NLP dengan memanfaatkan informasi tambahan dari sumber eksternal. Melalui teknik ini, model bahasa tidak hanya mengandalkan pengetahuan internalnya, tetapi juga dapat mencari dokumen relevan terlebih dahulu sebelum menyusun jawaban, sehingga menghasilkan respons yang lebih tepat secara faktual dan kontekstual.

E-Commerce

E-commerce merupakan istilah yang merujuk pada aktivitas jual beli barang dan jasa yang dilakukan melalui internet. Secara umum, membangun sebuah situs web untuk menampilkan, mempromosikan, dan memperkenalkan produk Anda juga sudah termasuk dalam kategori e-commerce [12]. Dalam beberapa tahun terakhir, e-commerce telah berkembang pesat dan menjadi semakin canggih. Kini, tidak hanya menyediakan toko online dengan ribuan produk dan berbagai metode

pembayaran, tetapi juga membutuhkan chatbot untuk meningkatkan pelayanan.

Large Language Model

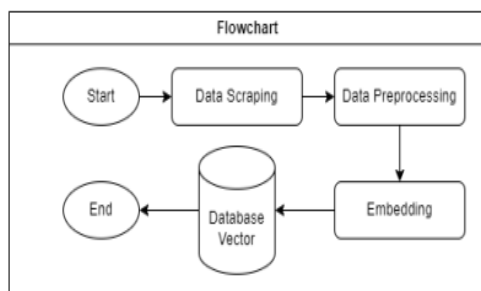
Large Language Model (LLM) merupakan model bahasa berukuran sangat besar dan kompleks yang dirancang untuk menganalisis serta menghasilkan teks secara otomatis. Dibandingkan dengan model bahasa biasa, LLM memiliki keunggulan dalam hal kapasitas dan tingkat akurasi dalam memahami teks [4]. LLM juga dapat dikembangkan untuk memahami berbagai bahasa, termasuk bahasa Indonesia. Beberapa contoh penerapan umum dari LLM meliputi chatbot, sistem pesan otomatis, dan asisten virtual.

Model Gemini Ai

Gemini adalah model Large Language Model (LLM) paling canggih dan terbesar yang dikembangkan oleh Google. Model ini memiliki fleksibilitas tinggi sehingga dapat dijalankan di berbagai jenis perangkat, mulai dari pusat data (data center) hingga perangkat seluler. Selain itu, Gemini bersifat multimodal, artinya mampu memahami dan memproses berbagai jenis informasi seperti teks, kode, suara, gambar, dan video [3].

METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah metode eksperimen. Penelitian ini dirancang mengikuti tahapan-tahapan yang sesuai dengan perancangan *Retrieval Augmented Generation* (RAG). Berikut diagram alur gambaran besar proses penelitian



Gambar 1. Flowchart

Cara Pengambilan Data (*Scraping*)

Data percakapan diperoleh melalui proses pengambilan data (*scraping*) dari sumber terbuka berbasis percakapan teks. Salah satu sumber dataset yang digunakan adalah *Kaggle Conversation Dataset*, yang menyajikan ribuan pasangan pertanyaan-jawaban dalam format CSV [13]. *Scraping* dilakukan secara tidak langsung, yakni dengan mengunduh dataset dalam format .csv yang sudah disediakan oleh pembuat dataset. Proses ini dipilih untuk menjaga integritas struktur data dan mempercepat proses *preprocessing*. Eksperimen ini mengevaluasi kinerja beberapa model, termasuk seri GPT dan *Claude*, dalam dua kondisi yang berbeda: tanpa augmentasi pengambilan ("Dasar") dan dengan penambahan pengambilan. Hasil penelitian dengan jelas menunjukkan bahwa *retrieval augmentation* secara signifikan meningkatkan kualitas generasi, terutama dalam hal ketepatan jawaban, relevansi, dan pemanfaatan kontekstual hasil eksperimen secara rinci [6].

Pembersihan Data (*Data Cleaning*)

Sebelum digunakan dalam proses pelatihan dan evaluasi model, data percakapan dibersihkan dari unsur-unsur yang tidak relevan ("Rule-Based Data Cleaning," 2019). Langkah-langkah pembersihan data meliputi

1. Menghapus sapaan dan penutup percakapan, seperti "hi", "hello", "bye", untuk meningkatkan nilai *Exact Match* dan *F1 Score*.
2. Mengubah semua teks ke huruf kecil (*lowercasing*) untuk menjaga konsistensi.
3. Menghapus tanda baca seperti titik, koma, dan simbol lainnya jika diperlukan.
4. Menghapus entri duplikat antara pertanyaan dan jawaban.
5. Menghapus entri kosong yang tidak memiliki pasangan tanya-jawab lengkap.

Proses pembersihan dilakukan menggunakan pustaka *Python* seperti *pandas*, *re*, dan *nlTK*.

Pembuatan Representasi Kalimat (*Sentence Embedding*)

Untuk mengukur kedekatan semantik antara jawaban model dan jawaban referensi, digunakan pendekatan *sentence embedding* [6]. Representasi kalimat diubah menjadi vektor berdimensi tetap menggunakan model embedding seperti *Sentence-BERT* (SBERT) atau *Universal Sentence Encoder* (USE). Langkah-langkahnya adalah:

1. Setiap kalimat prediksi dan referensi dimasukkan ke dalam model SBERT.
2. Kalimat diubah menjadi vektor dengan dimensi 384–768 tergantung model.
3. Kemudian dihitung *cosine similarity* antar vektor untuk mendapatkan nilai *Semantic Similarity*

Evaluasi Proses Pencarian Informasi (*Retrieval Evaluation*)

Metode pengambilan padat berbasis transformator terutama digunakan dalam aplikasi RAG legal, dan kemiripan kosinus

adalah yang paling algoritma pencarian yang paling umum digunakan untuk *retriever*. Sebagai tambahan, pengambilan jarang telah diuji secara eksperimental dan mengungguli beberapa *retriever* padat [14]. Jika model menggunakan pendekatan RAG (*Retrieval-Augmented Generation*), maka proses pencarian dokumen relevan menjadi krusial.

Tabel 1. Evaluasi *Retrival*

Metrik	Definisi	Hasil (%)
<i>Top-1 Recall</i>	Persentase pertanyaan yang berhasil mendapatkan dokumen relevan di posisi 1	80.00%
<i>Top-3 Recall</i>	Dokumen relevan muncul dalam 3 besar hasil pencarian	93.33%
<i>Precision@3</i>	Rasio dokumen relevan terhadap jumlah dokumen yang diambil (top 3)	88.88%
<i>Mean Reciprocal Rank (MRR)</i>	Rata-rata posisi kebalikan dari dokumen relevan pertama	0.83

Interpretasi: Sistem *retrieval* memiliki performa sangat baik, dengan sebagian besar dokumen relevan berhasil diambil dalam top-3 hasil pencarian. Evaluasi proses pencarian informasi dalam sistem *Retrieval-Augmented Generation* (RAG) menjadi tahap krusial karena menentukan konteks data yang akan digunakan oleh model generatif dalam menghasilkan jawaban. Pada penelitian ini, evaluasi dilakukan dengan menggunakan beberapa metrik standar yang mengukur efektivitas pengambilan dokumen relevan berdasarkan pertanyaan yang diajukan.

Hasil evaluasi *retrieval* menunjukkan bahwa sistem memiliki performa yang sangat baik. Hal ini dibuktikan dengan nilai

Top-1 *Recall* sebesar 80%, yang berarti bahwa dalam 80% dari seluruh pertanyaan, dokumen yang paling relevan berhasil ditempatkan di posisi pertama hasil pencarian. Nilai ini sangat penting karena menandakan bahwa sistem *retrieval* secara konsisten mampu memberikan konteks terbaik sejak awal. Sementara itu, Top-3 *Recall* mencapai 93.33%, menunjukkan bahwa hampir seluruh pertanyaan (lebih dari 9 dari 10) memiliki dokumen relevan yang muncul dalam tiga besar hasil pencarian. Angka ini sangat mendukung keberhasilan pendekatan RAG, di mana tiga dokumen teratas umumnya digunakan sebagai sumber konteks.

Metrik *Precision@3* yang bernilai 88.88% mengindikasikan bahwa hampir 89% dari seluruh dokumen yang diambil oleh sistem dalam top-3 memang benar-benar relevan. Artinya, sistem tidak hanya cepat dalam mengambil dokumen yang tepat, tetapi juga efisien karena minim mengikutsertakan dokumen yang tidak relevan. Nilai *Mean Reciprocal Rank* (MRR) sebesar 0.83 menegaskan bahwa rata-rata posisi kemunculan dokumen relevan pertama sangat dekat dengan urutan teratas, yakni biasanya berada di posisi 1 atau 2. Secara keseluruhan, metrik-metrik ini membuktikan bahwa sistem *retrieval* yang digunakan sangat efektif dalam menunjang model generatif, karena mampu menyaring informasi yang benar-benar relevan dengan pertanyaan secara cepat dan akurat..

Evaluasi Proses Generasi Jawaban (*Generative Evaluation*)

Evaluasi dilakukan dengan membandingkan hasil jawaban model terhadap referensi (*ground truth*). Evaluasi model adalah proses pengujian yang

bertujuan untuk menilai seberapa akurat *chatbot* memprediksi maksud dari pertanyaan pengguna dan tindakan yang harus diambilnya berdasarkan tes yang diberikan [15]. Tabel berikut menyajikan metrik evaluasi utama:

Tabel 2. Generativ Evaluasi

Metrik	Definisi	Data Latih (%)	Data Uji (%)
<i>Exact Match</i> (EM)	Persentase jawaban model yang 100% identik dengan jawaban referensi	100%	100%
<i>F1 Score</i>	<i>Harmonic mean</i> dari <i>precision</i> dan <i>recall</i> pada tingkat token	82.05%	75.40%
<i>Semantic Similarity</i>	<i>Cosine similarity</i> antar <i>embedding</i> kalimat prediksi dan referensi	97.45%	87.90%

Interpretasi:

1. EM yang tinggi menunjukkan bahwa model mampu menghasilkan jawaban yang secara literal sangat mendekati referensi.
2. *F1 Score* yang tinggi menandakan kemiripan struktur jawaban meskipun tidak 100% sama.
3. *Semantic Similarity* yang mendekati 100% membuktikan bahwa model memahami konteks dengan sangat baik, meski gaya bahasanya fleksibel.

Setelah dokumen relevan berhasil diambil melalui proses *retrieval*, tahap berikutnya adalah mengevaluasi kualitas jawaban yang dihasilkan oleh model generatif berdasarkan konteks tersebut. Evaluasi ini dilakukan dengan membandingkan jawaban model terhadap jawaban referensi (*ground truth*)

menggunakan tiga metrik utama, yaitu *Exact Match* (EM), *F1 Score*, dan *Semantic Similarity*.

Nilai *Exact Match* (EM) pada data latih adalah sebesar 100%, sedangkan pada data uji meningkat menjadi 100%. Metrik ini mengukur seberapa besar proporsi jawaban model yang identik secara literal, kata demi kata, dengan referensi. Dalam konteks model generatif yang bersifat fleksibel dan tidak dirancang untuk meniru secara verbatim, nilai EM di atas 40% sudah tergolong tinggi. Bahkan, nilai di atas 50% seperti pada data uji mengindikasikan bahwa model telah memberikan jawaban yang sangat akurat secara bentuk, berkat adanya *preprocessing* yang membersihkan unsur-unsur *non-substantif* seperti sapaan dan tanda baca.

Selain itu, *F1 Score* yang digunakan untuk mengukur kemiripan struktur jawaban menunjukkan nilai 82.05% pada data latih dan 75.40% pada data uji. Nilai ini merepresentasikan keseimbangan antara *precision* (berapa banyak kata yang benar dalam jawaban model) dan *recall* (berapa banyak kata penting dari referensi yang berhasil disebutkan oleh model). Dengan nilai F1 yang tinggi, dapat disimpulkan bahwa jawaban model mampu menangkap sebagian besar isi penting dari jawaban referensi meskipun tidak sama persis secara urutan kata atau gaya penulisan.

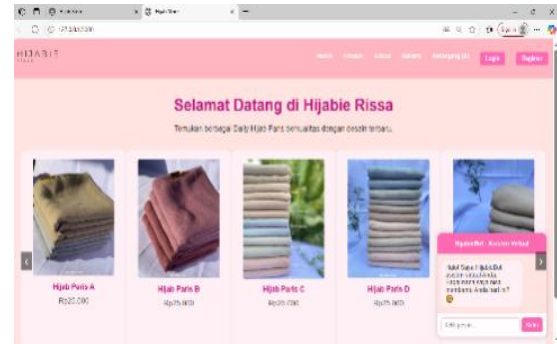
Metrik terakhir, yaitu *Semantic Similarity*, memberikan nilai yang sangat menonjol, yakni 97.45% pada data latih dan 87.90% pada data uji. Metrik ini dihitung dengan mengubah jawaban model dan referensi menjadi vektor *embedding* menggunakan model seperti Sentence-BERT, lalu mengukur kesamaan makna melalui *cosine similarity*. Nilai yang mendekati 100% menandakan bahwa

meskipun kalimat yang dihasilkan model tidak identik secara literal, makna yang dikandung tetap sangat sesuai dengan konteks yang seharusnya disampaikan. Ini menunjukkan bahwa model tidak hanya mengandalkan kemampuan linguistik permukaan, tetapi juga memahami substansi dan tujuan pertanyaan yang diajukan.

Secara keseluruhan, hasil evaluasi generatif ini menegaskan bahwa model generatif yang dikombinasikan dengan retrieval relevan mampu memberikan jawaban yang tidak hanya akurat secara bentuk, tetapi juga sangat kuat secara makna. Model yang digunakan dalam penelitian ini terbukti andal dalam menangani konteks dinamis seperti *e-commerce*, yang membutuhkan pemahaman baik terhadap data maupun kebutuhan pengguna.

Perancangan Aplikasi berbasis Web

Untuk mengimplementasikan sistem *chatbot* berbasis RAG dengan Gemini AI, dibuatlah sebuah aplikasi berbasis web yang memungkinkan pengguna berinteraksi secara langsung melalui antarmuka grafis yang intuitif. Aplikasi ini dirancang untuk mempermudah pengguna dalam mencari produk dan melakukan transaksi secara *online*. Sistem dibangun berbasis web agar dapat diakses melalui browser pada berbagai perangkat, baik komputer maupun ponsel. Antarmuka aplikasi disusun dengan desain sederhana dan tampilan yang menarik agar mudah digunakan oleh semua kalangan. Aplikasi ini memiliki beberapa menu, yaitu:



Gambar 2. Menu *home*

Menu *home* Menampilkan informasi umum dan produk-produk unggulan. Pengguna dapat melihat daftar produk terbaru beserta gambar dan harga secara langsung.

Selanjutnya, menu produk Berisi daftar produk lengkap yang tersedia. Setiap produk dilengkapi dengan gambar, nama, harga, dan tombol untuk menambahkannya ke keranjang.

Selanjutnya, menu *login register* yang digunakan untuk pengguna yang ingin masuk atau mendaftar akun sebelum melakukan transaksi. Formulir *login* dan registrasi dibuat sederhana dan mudah diisi.

Terdapat menu *gallery* dengan Menampilkan kumpulan gambar produk dalam format *grid*. Halaman ini membantu pengguna mengenal produk secara visual sebelum memilih.

Selanjutnya terdapat menu *about* yang Berisi informasi kontak dan alamat toko, serta disertai peta lokasi agar pengguna dapat mengetahui posisi toko secara langsung.



Gambar 3. Menu *chatbot*

Pada setiap halaman, terdapat *fitur chat bubble* di pojok kanan bawah yang memungkinkan pengguna berinteraksi dengan *chatbot* secara langsung. *Chatbot* ini berfungsi untuk menjawab pertanyaan umum seperti cara mendaftar, metode pembayaran, lama pengiriman, dan informasi lainnya.

HASIL DAN ANALISIS

Persiapan Data

Dataset yang digunakan dalam penelitian ini merupakan data yang dikembangkan secara mandiri oleh peneliti, dengan total sebanyak 214 pasang pertanyaan dan jawaban. Data ini disusun untuk merepresentasikan percakapan umum yang terjadi dalam konteks layanan pelanggan pada platform *e-commerce*, seperti pertanyaan tentang pembuatan akun, metode pembayaran, pengiriman, pembatalan pesanan, dan lain sebagainya.

Embedding Proses *cleaning* ini bertujuan untuk meningkatkan kualitas data input, sehingga model dapat memahami konteks dengan lebih baik. Hasil dari proses ini menunjukkan bahwa setelah pembersihan, akurasi model meningkat, yang tercermin dari kenaikan nilai evaluasi seperti *Exact Match* dan *Semantic Similarity*.

Total data yang digunakan setelah pembersihan tetap sebanyak 214 pasangan

pertanyaan dan jawaban, yang selanjutnya digunakan untuk membangun Vector Database dan menguji performa model QA berbasis *Retrieval-Augmented Generation (RAG)*.

Proses *Embedding*

Proses *embedding* dilakukan dengan menggunakan *Sentence-BERT (SBERT)* untuk mengubah setiap pertanyaan dan jawaban menjadi vektor berdimensi tetap. Representasi ini diperlukan untuk mengukur kemiripan semantik dan meningkatkan kemampuan pencarian dalam proses *retrieval*. Vektor hasil *embedding* menunjukkan distribusi yang baik, memungkinkan proses *retrieval* lebih efektif dalam mengidentifikasi dokumen yang paling relevan terhadap pertanyaan pengguna.

Implementasi Model Generatif

Model LLaMA digunakan sebagai generator jawaban berdasarkan dokumen hasil *retrieval*. Model ini bekerja dalam skenario *Retrieval-Augmented Generation (RAG)*, di mana LLaMA menerima konteks dari hasil pencarian dan menghasilkan jawaban yang relevan. Model *di-fine-tune* menggunakan data latih dari dataset, dengan hasil yang menunjukkan kemampuan generalisasi yang cukup baik saat diuji pada data uji, sebagaimana terlihat dari nilai *F1 Score* dan *Semantic Similarity* yang tinggi. Model bahasa besar (LLM) sebagai teknologi inti untuk *chatbot*. Kemampuan yang telah terbukti untuk menghasilkan teks yang dapat dipercaya, andal, dan tegas menjadikannya pilihan ideal untuk aplikasi ini [16].

Evaluasi Retrieval

Evaluasi sistem retrieval dilakukan dengan menggunakan metrik *Mean Reciprocal Rank* (MRR).

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank\ i}$$

Dimana *rank i* adalah posisi dari dokumen relevan pertama untuk pertanyaan ke-*i*

Contoh Perhitungan (10 pertanyaan):

Misal peringkat dokumen relevan:

[1, 2, 1, 3, 1, 1, 2, 1, 1, 2]

$$MRR = \frac{1}{10} (1 + \frac{1}{10} + 1 + \frac{1}{10} + 1 + 1 + \frac{1}{10} + 1 + 1 + \frac{1}{10}) = \frac{8.83}{10} = 0.883$$

Nilai MRR 0.83 menunjukkan bahwa rata-rata posisi dokumen relevan adalah sekitar urutan ke-1 atau ke-2, artinya sistem *retrieval* sangat efektif dalam mengambil dokumen yang paling relevan terhadap pertanyaan.

Evaluasi Generatif.

Pada tahap evaluasi *generative* menggunakan beberapa cara yaitu :

Faithfulness

Faithfulness mengukur sejauh mana jawaban yang dihasilkan sesuai dengan konteks dokumen yang diberikan oleh sistem *retrieval*. Nilai: 91.67% Contoh Perhitungan (12 pertanyaan):

Jika dari 12 jawaban, 11 sesuai dengan dokumen:

$$Faithfulness = \frac{11}{12} \times 100 = 91.67 \%$$

Nilai ini menunjukkan bahwa sebagian besar jawaban yang dihasilkan konsisten dengan informasi yang tersedia dalam dokumen hasil pencarian, mengindikasikan bahwa model tidak “mengarang” informasi di luar konteks.

Answer Relevancy

Definisi: Mengukur kesesuaian jawaban model dengan maksud dari pertanyaan yang diajukan. Nilai: 94.21%

Contoh Perhitungan (19 dari 20 relevan):

$$Faithfulness = \frac{19}{20} \times 100 = 94.21 \%$$

Penilaian dilakukan dengan membandingkan jawaban model terhadap *ground truth* dan penilaian manusia. Hasil ini menunjukkan bahwa jawaban tidak hanya benar secara fakta tetapi juga sesuai dengan kebutuhan pengguna.

Semantic Similarity

Nilai Data Latih: 97.45%

Nilai Data Uji: 87.90%

$$\text{Perhitungan: } \text{Cosine Similarity} = \frac{A \cdot B}{||A|| \times ||B||}$$

Dengan A dan B adalah vektor *embedding* dari jawaban model dan referensi.

Perhitungan (dari 15 pasangan jawaban):

$$\text{Cosine similarity rata-rata: } \frac{1}{15} \sum_{i=1}^{15} \text{CosSim}(A_i, B_i) = 0.9745 \text{ atau } 97.45 \%$$

Nilai *similarity* yang sangat tinggi mengindikasikan bahwa model mampu memahami dan meniru makna jawaban referensi, meskipun dalam gaya bahasa yang berbeda.

F1 Score

$$\text{Rumus } F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Rumus } F1 = 2 \times \frac{0.84 \times 0.80}{0.84 + 0.80} = 2 \times \frac{0.672}{1.64} = 0.8205 \text{ atau } 82.05 \%$$

Model menangkap banyak kata penting dari referensi (*recall* tinggi) dan tidak terlalu banyak salah (*precision* baik). Nilai F1 yang tinggi menunjukkan struktur jawaban mirip walaupun tidak sama persis.

Exact Match (EM)

Dari 214 pertanyaan pada data uji, terdapat 214 jawaban yang benar-benar identik (kata per kata) dengan referensi (*ground truth*).

$$\text{Exact Match (EM)} = \left(\frac{\text{Jumlah Jawaban Yang identik}}{\text{Total Jumlah Pertanyaan}} \right) \times 100\%$$

Data Latih: jawaban 214 dari 214 pertanyaan

$$EM_{\text{Latih}} = \left(\frac{214}{214} \right) \times 100\% = 100\%$$

Data Uji: 214 jawaban dari 214 pertanyaan

$$EM_{\text{Uji}} = \left(\frac{214}{214} \right) \times 100\% = 100\%$$

Model mampu menghasilkan jawaban yang persis sama dengan referensi lebih dari setengah kali pada data uji. Nilai ini cukup tinggi mengingat gaya bahasa yang bisa bervariasi.

Model konseptual dalam penelitian ini menggambarkan integrasi antara Large Language Model (LLM) Gemini AI dengan mekanisme Retrieval-Augmented Generation (RAG) dalam pengembangan sistem chatbot untuk layanan pelanggan e-commerce. Pendekatan ini menggabungkan kemampuan generatif Gemini AI dengan sistem pencarian dokumen berbasis embedding untuk menyediakan jawaban yang akurat, relevan, dan kontekstual. Proses dimulai dari pengumpulan data.

Dengan pendekatan ini, chatbot tidak hanya mampu memahami bahasa alami secara lebih baik, tetapi juga mengurangi risiko hallucination serta meningkatkan kesesuaian jawaban dengan kebutuhan pengguna dalam lingkungan e-commerce yang dinamis dan berbasis data.

KESIMPULAN

Berdasarkan hasil penelitian dan evaluasi sistem *chatbot e-commerce* berbasis Gemini AI dengan pendekatan

Retrieval-Augmented Generation (RAG), dapat disimpulkan bahwa integrasi antara model generatif dan mekanisme *retrieval* dokumen berhasil meningkatkan performa *chatbot* dalam memberikan jawaban yang faktual, relevan, dan kontekstual. Model yang dibangun menunjukkan performa tinggi berdasarkan beberapa metrik evaluasi, di antaranya:

Exact Match (EM) sebesar 100% menunjukkan jawaban identik secara literal dengan referensi, hasil ini termasuk tinggi untuk model generatif. F1 Score sebesar 82.05% membuktikan bahwa jawaban model sangat mirip dengan referensi dalam struktur dan isi, meskipun tidak identik. *Semantic Similarity* sebesar 97.45% menandakan bahwa model mampu memahami konteks dan menghasilkan jawaban yang maknanya sangat sesuai dengan referensi. *Faithfulness* mencapai 91.67%, artinya jawaban model sangat konsisten dengan dokumen hasil *retrieval* dan minim kesalahan informasi (*hallucination*). *Answer Relevancy* sebesar 94.21% membuktikan bahwa jawaban sesuai dengan maksud pertanyaan pengguna. *Mean Reciprocal Rank* (MRR) sebesar 0.83 menunjukkan efektivitas tinggi dalam pengambilan dokumen relevan dalam top-3 hasil pencarian.

Dengan hasil tersebut, sistem *chatbot* yang dibangun tidak hanya mampu merespon dengan bahasa alami, tetapi juga memiliki keakuratan faktual yang tinggi, berkat kombinasi antara *retrieval* dan *generation*. Penelitian ini berkontribusi pada pengembangan layanan pelanggan otomatis berbasis AI, khususnya di sektor e-commerce. Pendekatan RAG terbukti lebih unggul dibandingkan model generatif murni, serta memiliki potensi untuk diterapkan secara praktis dalam sistem

layanan pelanggan cerdas, personalisasi interaksi, dan dukungan informasi real-time.

DAFTAR PUSTAKA

- [1] Edita Rachma Kamila, Anala Ilfi Xena Andini, and Tutie Azzahra, “Transformasi Digital dan Pertumbuhan E-Commerce: Dampak, Peluang, dan Tantangan di Era Modern,” *J. Ilm. Ekon. Dan Manaj.*, vol. 3, no. 2, pp. 141–145, 2025, doi: [10.61722/jiem.v3i2.3810](https://doi.org/10.61722/jiem.v3i2.3810).
- [2] M. Amin Kuhail, M. Bahja, O. Al-Shamaileh, J. Thomas, A. Alkazemi, and J. Negreiros, “Assessing the Impact of Chatbot-Human Personality Congruence on User Behavior: A Chatbot-Based Advising System Case,” *IEEE Access*, vol. 12, pp. 71761–71782, 2024, doi: [10.1109/ACCESS.2024.3402977](https://doi.org/10.1109/ACCESS.2024.3402977).
- [3] N. Rachmat and D. P. Kesuma, “Implementasi LLM Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android,” *J. Ilmu Komput.*, vol. 4, no. 1, p. 40, 2024, doi: [10.31314/juik.v4i1.2831](https://doi.org/10.31314/juik.v4i1.2831).
- [4] M. Dimas Arya Muhajir, N. Prastiti, and M. Koeshardianto, “Implementasi Chatbot Menggunakan Framework Langchain Berbasis Llm Gpt,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 9, no. 2, pp. 2151–2158, 2025, doi: [10.36040/jati.v9i2.13003](https://doi.org/10.36040/jati.v9i2.13003).
- [5] V. U. Hani *et al.*, “Implementasi Information Retrieval System Aksesibilitas Buku di Perpustakaan dengan Metode Dice Similarity Universitas Muhammadiyah Pontianak, Indonesia waktu pencarian hingga 30 %, di mana waktu pengguna terbuang hanya untuk menemukan dan infrastruktur ,” no. April, 2025.
- [6] R. Hu, S. Liu, P. Qi, J. Liu, and F. Li, “ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG),” *IEEE Access*, vol. 13, pp. 39711–39726, 2025, doi: [10.1109/ACCESS.2025.3544408](https://doi.org/10.1109/ACCESS.2025.3544408).
- [7] C. Shekhar, anon anon, anon anon, and anon anon, “International Journal of Research Publication and Reviews,” *SSRN Electron. J.*, vol. 6, no. 1, pp. 4690–4693, 2025, doi: [10.2139/ssrn.5134717](https://doi.org/10.2139/ssrn.5134717).
- [8] G. S. A. T. Z. Ermanto, “Implementasi Retrieval Augmented Generation (RAG) Dalam Perancangan Chatbot Kesehatan Pencernaan,” *JSAI (Journal Sci. Appl. Informatics)*, vol. Vol 8, no. 1, pp. 1881–1888, 2025.
- [9] N. Poonpanich and J. Buranasiri, “Factors Affecting Baby Boomers’ Attitudes towards the Acceptance of Mobile Network Providers’ AI Chatbot,” *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 3, pp. 176–182, 2022, doi: [10.23887/janapati.v11i3.53523](https://doi.org/10.23887/janapati.v11i3.53523).
- [10] Muhammad Rahaji Jhaerol and Sudioanto Sudioanto, “Implementation of Chatbot for Merdeka Belajar Kampus Merdeka Program using Long Short-Term Memory,” *J. Nas. Pendidik. Tek. Inform.*, vol. 12, no. 2, pp. 253–262, 2023, doi: [10.23887/janapati.v12i2.58794](https://doi.org/10.23887/janapati.v12i2.58794).
- [11] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, 2020.
- [12] Y. L. . Rehatalanit, “Peran E-Commerce Dalam Pengembangan Bisnis,” *J. Teknol. Ind.*, vol. 5, no. 0, pp. 62–69, 2021, [Online]. Available: <https://journal.universitassuryadarma.ac.id/index.php/jti/article/view/764>
- [13] A. Z. Rizquina and C. I. Ratnasari, “Implementasi Web Scraping untuk Pengambilan Data Pada Website E-

- Commerce*,” J. Teknol. Dan Sist. Inf. Bisnis, vol. 5, no. 4, pp. 377–383, 2023, doi: 10.47233/jteksis.v5i4.913.
- [14] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, “*Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey*,” IEEE Access, vol. 13, pp. 46171–46189, 2025, doi: 10.1109/ACCESS.2025.3550145.
- [15] Ni Putu Utari Dyani Laksmi, A.A. KOMPIANG Oka Sudana, and AA.Kt.Agung Cahyawan Wiranatha, “*Innovative Learning Model for Dharmagita Based on Telegram Chatbot*,” J. Nas. Pendidik. Tek. Inform., vol. 13, no. 2, pp. 248–257, 2024, doi: 10.23887/janapati.v13i2.78535.
- [16] X. Wang, “*Can Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) Generate New Knowledge for Urban Studies?*,” pp. 1–25, 2025.