PENGEMBANGAN CHATBOT E-COMMERCE BERBASIS LARGE LANGUAGE MODEL DENGAN PENDEKATAN RETRIEVAL-AUGMENTED GENERATION UNTUK MENDUKUNG AUTOMATED QUERY RESOLUTION DAN ORDER PROCESSING

Muhammad Haris Sitompul[1], Nur Ichsan Utama[2], Sinung Suakanto[3]

1st Muhammad Haris Sitompul
Fakultas Rekayasa Industri
Telkom University
Bandung, Indonesia
muhariss@student.telkomuniversity.ac.id

2nd Nur Ichsan Utama
Fakultas Rekayasa Industri
Telkom University
Bandung, Indonesia
nichsan@telkomuniversity.ac.id

3rd Sinung Suakanto
Fakultas Rekayasa Industri
Telkom University
Bandung, Indonesia
sinung@telkomuniversity.ac.id

Abstrak*— Seiring dengan pesatnya perkembangan teknologi informasi dan komunikasi di era digital, lanskap bisnis global mengalami perubahan yang signifikan, salah satunya transformasi digital melalui platform e-commerce. Namun, seiring meningkatnya jumlah produk dan kebutuhan informasi yang akurat bagi pelanggan, tantangan muncul dalam menyediakan layanan pelanggan yang responsif dan relevan. Penelitian ini mengusulkan pengembangan chatbot e-commerce dengan pendekatan Retrieval-Augmented Generation (RAG), yang menggabungkan kemampuan Large Language Model (LLM) dengan sistem pencarian dokumen berbasis vektor. Model LLM yang digunakan adalah Llama-3.3-70B-Instruct, yang telah ditingkatkan kemampuannya dengan menambahkan informasi relevan melalui pencarian semantik terhadap knowledge base yang disimpan dalam vector storage berupa FAISS. Dengan pendekatan ini, chatbot mampu memberikan jawaban berbasis data aktual tanpa perlu melakukan fine-tuning, serta meminimalkan munculnya jawaban yang bersifat asumsi atau spekulatif. Hasil implementasi sistem menunjukkan bahwa integrasi LLM dan RAG dapat meningkatkan efisiensi layanan pelanggan dalam platform e-commerce. Hal ini dibuktikan melalui evaluasi mengunakan performance metrics dengan hasil skor metrik yang cukup tinggi, sehingga menunjukkan bahwa chatbot mampu memberikan jawaban yang akurat dan relevan sesuai kebutuhan pengguna.

***Kata kunci*— e-commerce, chatbot, Large Language Model, Retrieval-Augmented Generation, LLaMA, FAISS**

I. PENDAHULUAN

Dalam era globalisasi dan perkembangan teknologi informasi, metode bisnis mengalami perubahan yang signifikan. Salah satu inovasi dari perkembangan teknologi dalam dunia bisnis adalah *e-commerce*. Sektor *e-commerce* telah menjadi salah satu pendorong utama dalam perekonomian dunia. Berdasarkan data yang dipublikasikan oleh Badan Pusat Statistik (BPS) pada tahun 2023, pertumbuhan volume transaksi *online* telah meningkat pesat setiap tahunnya, terutama didorong oleh perubahan perilaku konsumen yang semakin terbiasa berbelanja secara digital. Hal ini mencerminkan semakin meningkatnya preferensi konsumen terhadap belanja daring melalui *e-commerce*. Proyeksi pendapatan *e-commerce* diperkirakan akan mencapai US$4,18 triliun pada tahun 2024 [1]. Data-data tersebut menunjukkan betapa besarnya pengaruh *e-commerce* dalam transformasi dunia bisnis, serta dampaknya yang semakin kuat terhadap ekonomi global di masa depan [1].

Seiring bertambahnya jumlah produk yang ditawarkan di platform *e-commerce*, kompleksitas pengelolaan data produk juga meningkat secara signifikan. Kondisi ini menciptakan tantangan besar bagi penyedia layanan dalam memastikan bahwa informasi yang disampaikan kepada pelanggan tetap akurat, konsisten, dan selalu diperbarui sesuai perubahan data terbaru. Permasalahan semakin rumit karena deskripsi produk yang panjang, format yang beragam, serta perbedaan istilah antara sistem dan bahasa yang digunakan oleh pelanggan. Sementara itu, konsumen *modern* mengharapkan informasi yang detail, ringkas, dan mudah dipahami untuk mempercepat pengambilan keputusan, sehingga diperlukan sistem yang mampu mengelola volume data besar sekaligus memahami kebutuhan pengguna secara kontekstual.

Untuk menjawab tantangan tersebut, penelitian ini mengusulkan pengembangan *chatbot e-commerce* dengan pendekatan RAG. Sistem ini dirancang agar mampu memberikan jawaban yang lebih akurat dan kontekstual dengan memanfaatkan kemampuan LLM yang diperkuat oleh proses pencarian dokumen berbasis vektor. Melalui pendekatan ini, *chatbot* tidak hanya mengandalkan pengetahuan internal, tetapi juga dapat merespons pertanyaan pengguna dengan data produk terkini yang diperoleh secara dinamis dari *knowledge base*. Implementasi dan evaluasi sistem juga dilakukan untuk memastikan efektivitas dalam meningkatkan kualitas layanan informasi produk, relevansi jawaban, dan kepuasan pengguna di platform *e-commerce*.

II. KAJIAN TEORI

Untuk mendukung pengembangan sistem *chatbot e-commerce* yang responsif dan berbasis data aktual, diperlukan pemahaman mendalam mengenai teknologi LLM serta pendekatan *retrieval* yang mampu menangani kompleksitas data produk secara kontekstual.

A. Chatbot

Chatbot merupakan program komputer yang dirancang untuk merespons pertanyaan pengguna secara otomatis dengan pengalaman interaksi menyerupai percakapan manusia, baik melalui teks, suara, maupun elemen visual [2]. Perkembangannya terbagi menjadi *chatbot* berbasis *rule-based* yang lebih terstruktur dan prediktif [3], serta *chatbot*

modern berbasis LLM yang lebih fleksibel menangani konteks percakapan kompleks [4]. Dalam konteks *e-commerce*, *chatbot* tidak hanya meningkatkan efisiensi layanan dan memberikan respons 24 jam [2], tetapi juga berperan sebagai asisten penjualan cerdas yang mampu memberikan rekomendasi produk dan memengaruhi keputusan pembelian pelanggan [5], [6]. Integrasi *chatbot* terbukti membantu meningkatkan kepuasan, loyalitas, dan efektivitas transaksi, bahkan dalam lingkungan terbatas seperti pusat perbelanjaan kampus [7].

B. Large Language Model (LLM)

LLM memanfaatkan teknologi berbasis *Natural Language Processing* (NLP) untuk memahami, menafsirkan, dan menghasilkan teks dalam jumlah besar secara kontekstual, yang memungkinkan otomatisasi layanan, analisis sentimen, dan rekomendasi produk yang lebih relevan di *e-commerce* [1], [8]. Dengan teknik *fine-tuning* dan personalisasi, LLM dapat meningkatkan akurasi respons dan mempercepat waktu tanggap, sehingga mendukung interaksi *real-time* yang lebih adaptif dan meningkatkan kepuasan pelanggan [8].

C. Retrieval-Augmented Generation (RAG)

RAG adalah pendekatan yang menggabungkan *retrieval* berbasis pencarian semantik dan *generation* berbasis LLM untuk menghasilkan jawaban yang lebih faktual, relevan, dan kontekstual [9]. Dengan memanfaatkan dokumen eksternal yang selalu diperbarui, RAG membantu mengatasi keterbatasan model generatif murni, mengurangi risiko halusinasi, serta memungkinkan pembaruan konten tanpa *retraining* penuh, sehingga cocok digunakan di domain *e-commerce* dan pendidikan yang memerlukan informasi terkini dan akurat [10].

D. Vector Store

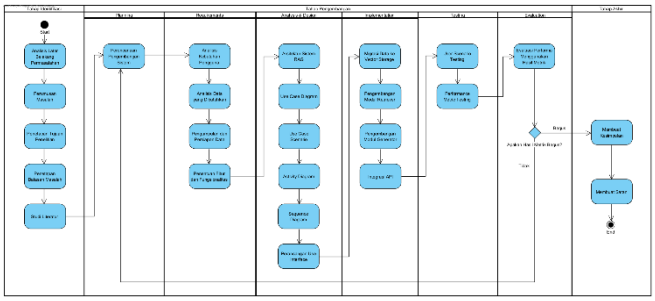
Vector store adalah sistem untuk menyimpan dan mencari *embedding* berdimensi tinggi, yang merepresentasikan kesamaan makna data seperti teks atau gambar [11]. Teknologi seperti FAISS memungkinkan pencarian kemiripan vektor secara efisien melalui algoritma *Approximate Nearest Neighbor Search*, sehingga mendukung *retrieval* data relevan dalam skala besar tanpa hanya bergantung pada pencocokan kata kunci literal [12]. Pendekatan ini sangat penting dalam arsitektur RAG karena mempercepat pencarian konteks yang sesuai untuk menjawab pertanyaan pengguna secara lebih akurat.

III. METODE

Dalam menyelesaikan masalah pada penelitian ini, metode penyelesaian yang digunakan adalah pendekatan pengembangan sistem berbasis SDLC (*Software Development Life Cycle*) yaitu metode *Iterative Incremental Model*. Pendekatan ini diharapkan dapat mendukung proses pembangunan *chatbot* yang responsif, faktual, serta relevan terhadap kebutuhan pengguna.

A. Sistematika Penyelsaian Masalah

Penyelesaian masalah dalam penelitian ini dilakukan secara bertahap, dimulai dari proses identifikasi untuk memahami dan merumuskan permasalahan yang ada. Selanjutnya, dilakukan tahap perencanaan dan analisis kebutuhan untuk menentukan arah pengembangan sistem yang sesuai dengan tujuan penelitian. Hasil analisis ini kemudian diwujudkan ke dalam desain arsitektur dan diagram untuk memetakan alur kerja sistem secara lebih terstruktur. Tahap berikutnya adalah implementasi, di mana sistem yang telah dirancang direalisasikan ke dalam bentuk fungsional. Setelah sistem berhasil diimplementasikan, sistem diuji untuk memastikan kinerjanya sesuai dengan harapan melalui pengujian berbasis metrik yang relevan. Hasil metrik tersebut kemudian digunakan dalam tahap evaluasi untuk menilai efektivitas dan kualitas sistem, sekaligus menjadi dasar perbaikan dan penyempurnaan agar sistem dapat memberikan hasil yang lebih optimal. Pendekatan sistematis ini diharapkan mampu menghasilkan solusi yang tidak hanya relevan dengan kebutuhan pengguna, tetapi juga adaptif terhadap perkembangan data dan teknologi.



Gambar II- 1 Sistematika Penyelesaian Masalah

B. Metode Pengumpulan Data

Metode pengumpulan data dalam penelitian ini dilakukan melalui kombinasi antara metode observasi langsung terhadap sistem atau proses *existing* pada platform *e-commerce*, wawancara dengan *stakeholder*, serta studi literatur agar dapat menangkap kebutuhan sistem secara holistik dan mendalam. Dengan pendekatan ini dapat membantu peneliti dalam memahami ekspektasi pengguna yang menjadi dasar perancangan sistem *chatbot* RAG. Data yang dikumpulkan dalam penelitian ini terdiri dari dua jenis utama yang menjadi fondasi bagi pengembangan dan evaluasi *chatbot* RAG. Pertama, dilakukan pengumpulan data untuk membangun *knowledge base* yang akan digunakan oleh sistem *retrieval*. Data ini merupakan data primer yang bersumber dari katalog produk dalam platform *e-commerce*, mencakup informasi terstruktur seperti nama produk, deskripsi detail, spesifikasi teknis, harga, kategori, brand, ketersediaan stok, dan data lainnya. Data katalog ini diekstraksi dan dibersihkan untuk memastikan kualitas dan konsistensi informasi sebelum diindeks ke dalam *knowledge base*. Kedua, untuk keperluan pengujian dan evaluasi performa *chatbot*, dibuat sebuah dataset tanya-jawab (*question-answering*). Dataset ini disusun secara manual dengan merumuskan berbagai macam pertanyaan yang merepresentasikan kueri pengguna di dunia nyata, mulai dari pertanyaan umum tentang fitur produk hingga meminta rekomendasi sesuai dengan kriteria atau keinginan pengguna.

C. Metode Pengolahan Data

Berikut adalah tahapan metode pengolahan data yang dilakukan untuk mempersiapkan data produk menjadi basis pengetahuan yang siap digunakan oleh sistem *chatbot* berbasis RAG

1. Sumber dan Ekstraksi Data

Data diperoleh langsung dari *database* platform *e-commerce* menggunakan *query* SQL yang hanya mengambil data produk relevan, sehingga selalu sesuai kondisi terbaru dan langsung siap diproses lebih lanjut.

2. Seleksi Atribut

Dilakukan penyaringan terhadap atribut-atribut penting yang relevan bagi pengguna *chatbot*. Atribut yang dipilih terutama difokuskan pada informasi yang paling sering dicari dan dibutuhkan oleh pengguna, seperti nama produk, deskripsi detail, kategori, *brand*, harga, ketersediaan, dan poin-poin utama spesifikasi atau fitur penting lainnya

3. *Data Preprocessing*

Data yang dipilih menjalani tahapan normalisasi, termasuk mengubah semua teks menjadi huruf kecil (*lowercasing*), menghapus karakter tidak relevan (*noise*), dan menghilangkan elemen-elemen HTML atau simbol non-alfabet yang dapat mengganggu proses *embedding*.

4. Strukturisasi Dokumen

Setiap entri produk disusun ulang menjadi satu blok teks utuh yang menggabungkan seluruh atribut penting menjadi satu dokumen yang koheren. Struktur dokumen dirancang secara terstuktur agar setiap dokumen tidak hanya berisi data mentah, tetapi juga menghadirkan konteks yang koheren, ringkas, dan informatif, sehingga dapat mendukung kinerja optimal sistem *chatbot* dalam memenuhi kebutuhan pengguna.

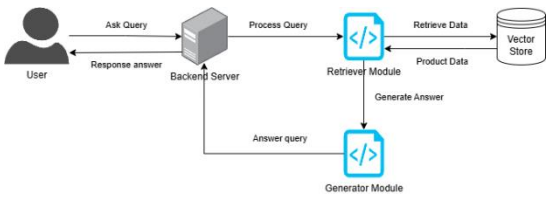
5. *Embedding* Produl

Dokumen yang sudah dihasilkan diubah menjadi vektor numerik menggunakan *embedding* model, lalu disimpan kedalam *vector store* berupa FAISS, memungkinkan pencarian berbasis kemiripan makna (*semantic search*) untuk mendukung performa *chatbot* berbasis RAG.

IV. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah implementasi sistem *chatbot* berbasis *Retrieval-Augmented Generation* (RAG) pada domain *e-commerce*, serta membahas performa sistem berdasarkan pengujian dan evaluasi yang telah dilakukan. Penjabaran mencakup kinerja *retrieval* dokumen, kualitas jawaban yang dihasilkan oleh model generatif, serta analisis terhadap efektivitas sistem dalam merespons pertanyaan pengguna secara relevan dan berbasis fakta.

A. Arsitektur Sistem



Gambar IV- 1 Arsitektur Sistem RAG

Arsitektur sistem yang diimplementasikan dalam penelitian ini menggabungkan komponen *retrieval* dan generatif dalam satu alur terpadu berbasis pendekatan RAG. Proses dimulai ketika pengguna mengajukan pertanyaan melalui antarmuka, baik itu berupa halaman web, *chatbot* API, atau kanal digital lainnya. Pertanyaan ini dikirim ke *backend service* yang bertanggung jawab untuk memproses input dan melakukan validasi. Pertanyaan kemudian diproses melalui model *embedding* BAAI/bge-m3 untuk diubah menjadi representasi vektor berdimensi tetap. Representasi vektor ini digunakan sebagai kunci pencarian terhadap koleksi dokumen produk yang telah di-*embedding* dan disimpan di dalam *vector store* berbasis FAISS.

Modul *retriever* melakukan pencarian semantik berdasarkan kedekatan vektor menggunakan algoritma *Approximate Nearest Neighbor* (ANN) untuk mengambil top-k dokumen produk yang paling relevan. Dokumen-dokumen ini digabungkan bersama pertanyaan pengguna dan diberikan sebagai input ke modul *generator*, yang berisi model LLM dengan versi Llama-3.3-70B-Instruct. Di dalam modul *generator*, *prompt* disusun secara khusus untuk mengatur gaya bahasa, struktur, dan format jawaban agar sesuai dengan konteks *e-commerce*, sehingga LLM dapat langsung menghasilkan respons yang informatif, ringkas, dan konsisten dengan kebutuhan pengguna. Jawaban yang dihasilkan dikirimkan kembali ke antarmuka pengguna. Dengan arsitektur ini, sistem tidak hanya mengandalkan memori internal LLM, tetapi juga mampu merespons pertanyaan spesifik berbasis data aktual dan relevan yang terus diperbarui di *knowledge base*.

B. Komponen *Retriever*

Komponen *Retriever* merupakan inti dari pendekatan RAG yang digunakan dalam sistem *chatbot* ini. Tugas utamanya adalah menjembatani pertanyaan pengguna yang bersifat bebas dan beragam dalam bahasa natural dengan basis pengetahuan terstruktur berupa representasi vektor dari dokumen produk. Tujuan utama modul *retriever* adalah memastikan bahwa informasi yang paling relevan dapat ditemukan dan diambil dari *kmowledge base* secara cepat dan akurat. Kualitas dokumen hasil pencarian ini sangat berpengaruh terhadap konteks yang akan diberikan ke LLM, sehingga akan berdampak langsung pada akurasi jawaban yang dihasilkan. Untuk itu, sistem ini menggunakan pendekatan *semantic search* berbasis representasi vektor, yang dinilai lebih unggul dibanding pencarian berbasis kata kunci (*keyword matching*) dalam memahami makna kalimat.

Proses *retrieval* dimulai dengan melakukan transformasi query, yakni menambahkan awalan seperti “query:” pada pertanyaan pengguna agar model *embedding* dapat membedakan antara teks kueri dan dokumen. Selain itu, ditambahkan juga *instruction* tambahan yang berfungsi sebagai petunjuk bagi komponen *retriever* untuk memperjelas konteks pencarian. Setelah itu, kueri diubah menjadi vektor numerik menggunakan model *embedding* yang sama dengan yang digunakan saat fase indeksasi dokumen. Langkah ini bertujuan untuk membantu sistem dalam memahami maksud pertanyaan atau permintaan pengguna secara semantik. Vektor hasil *embedding* ini kemudian dicocokkan ke dalam basis data vektor (FAISS) menggunakan metrik *cosine similarity*, yang mengukur seberapa mirip arah antara dua vektor dalam ruang semantik. Hasil pencarian berupa daftar peringkat dokumen dengan skor kemiripan tertinggi, yang kemudian diteruskan ke modul *generator* untuk membentuk jawaban berbasis konteks yang ditemukan.

C. Komponen *Generator*

Komponen *Generator* merupakan bagian hilir dari system RAG, yang berperan penting dalam menghasilkan jawaban akhir yang relevan dan koheren. Setelah dokumen relevan diperoleh oleh *retriever*, dokumen tersebut tidak langsung

diberikan kepada pengguna, melainkan dikombinasikan dengan pertanyaan asli dalam bentuk *prompt* terpadu. *Prompt* ini kemudian dikirimkan ke LLM untuk menghasilkan teks jawaban. Jika *retriever* berfungsi sebagai pencari informasi, maka *generator* bertindak sebagai penyaji informasi yang memahami konteks dan menyampaikannya dalam bentuk percakapan alami yang sesuai dengan kebutuhan pengguna. Pembuatan *prompt* dilakukan secara sistematis dengan menyertakan tiga elemen utama:

1. Instruksi Sistem (*System Instruction*)

Bagian ini mendefinisikan persona dan aturan ketat untuk LLM. Instruksi ini secara eksplisit memerintahkan LLM untuk bertindak sebagai asisten *e-commerce* yang membantu dan mendasarkan jawabannya hanya pada konteks yang diberikan. Dengan instruksi ini, model diharuskan mengabaikan bagian data yang tidak terkait dengan pertanyaan dan dilarang menambahkan asumsi atau pengetahuan eksternal yang tidak terdapat dalam konteks. Tujuannya adalah untuk menjaga agar jawaban yang dihasilkan tetap akurat, konsisten, dan fokus pada data yang benar-benar relevan

2. Penyajian Konteks (*Context Presentation*)

Data produk yang telah diambil oleh komponen *retriever* disajikan dalam format yang jelas dan terstruktur. Setiap dokumen produk diberi penanda agar LLM dapat dengan mudah membedakan sumber informasi yang berbeda. Teknik ini penting untuk membedakan antara berbagai sumber yang mungkin memiliki informasi serupa namun tidak identik

3. Use *Query*

Pertanyaan asli dari pengguna dicantumkan secara eksplisit di akhir *prompt*. Hal ini membantu LLM memahami dengan tepat apa yang harus dijawab, serta menjaga agar respons tetap fokus dan relevan dengan konteks dokumen yang ada. Penyajian *query* secara langsung juga berfungsi sebagai *instruction anchor*, yang menjaga agar respons tetap berada dalam ruang lingkup informasi yang relevan dan tidak menyimpang dari tujuan utama pertanyaan pengguna.

D. Komponen API

Komponen ini berfungsi sebagai penghubung antara antarmuka pengguna (*frontend*) dengan seluruh logika *backend* sistem *chatbot*, termasuk modul *retriever* dan *generator*. API dirancang menggunakan *framework* FastAPI, yang mendukung performa tinggi dan kemudahan dokumentasi otomatis. Setiap permintaan dari pengguna akan dikirimkan ke *endpoint* API dalam bentuk pertanyaan teks. *Endpoint* ini kemudian akan memproses input, memanggil *pipeline* RAG secara berurutan, dimulai dari *embedding* pertanyaan, pencarian dokumen dengan *retriever*, penyusunan *prompt*, hingga pemanggilan model LLM dan mengembalikan jawaban ke *end user*. Selain *endpoint* utama untuk memproses pertanyaan pengguna, tersedia juga *endpoint* khusus untuk admin yang memungkinkan pengelolaan system RAG, seperti memperbarui data *embedding*, menguji hasil *retrieval*, dan mengatur konfigurasi RAG secara dinamis. *Endpoint* untuk admin ini penting agar sistem dapat tetap relevan dan responsif terhadap perubahan data produk maupun kebutuhan bisnis, serta memudahkan pemeliharaan dan optimasi tanpa perlu mengganggu layanan yang berjalan. Dengan adanya komponen API, sistem *chatbot* dapat diintegrasikan dengan berbagai platform secara fleksibel dan menjaga alur komunikasi tetap efisien dan *real-time*.

E. Evaluasi Modul *Retriever*

Tabel IV- 1 Evaluasi Modul Retriever

Metrik Evaluasi	Nilai
Precision@k	0.3200
Recall@k	0.7867
Mean Reciprocal Rank (MRR)	0.8667

Evaluasi terhadap modul *retriever* dilakukan untuk mengukur kemampuan sistem dalam mengambil dokumen yang paling relevan terhadap pertanyaan pengguna sebelum diteruskan ke tahap generasi jawaban. Proses evaluasi ini mencakup pengujian pada beberapa *query* uji yang mewakili skenario nyata dalam penggunaan *chatbot e-commerce*.

Berdasarkan hasil evaluasi ketiga mertrik diatas, sistem *retrieval* menunjukkan performa yang cukup baik secara keseluruhan. Nilai metrik Precision@k sebesar 0.3200 mengindikasikan bahwa sekitar 32% dari dokumen yang ditampilkan di posisi teratas merupakan dokumen relevan, menandakan adanya ruang untuk peningkatan dalam menyaring hasil yang lebih tepat. Namun, Recall@k yang tinggi sebesar 0.7867 menunjukkan bahwa sebagian besar dokumen relevan berhasil ditemukan di antara hasil yang diambil, mencerminkan cakupan sistem yang cukup luas. Sementara itu, metrik *Mean Reciprocal Rank* (MRR) mencapai 0.8667, yang berarti bahwa dokumen relevan pertama rata-rata ditemukan pada posisi sangat awal dalam daftar hasil, mengindikasikan efektivitas tinggi dalam menghadirkan jawaban relevan secara cepat. Hasil performa *retrieval* menunjukkan bahwa sistem mampu menemukan sebagian besar dokumen relevan secara efektif dan menempatkannya pada posisi atas, sebagaimana ditunjukkan oleh nilai *recall* dan MRR yang tinggi. Nilai *precision* yang relatif rendah disebabkan oleh masih adanya dokumen tidak relevan yang ikut terambil dalam top-k, terutama ketika jumlah dokumen relevan per *query* terbatas.

F. Evaluasi Modul Generator

Tabel IV- 2 Evaluasi Modul Generator

Metrik Evaluasi	Nilai
BERT Score	0.7750
ROUGE-1	0.5032
ROUGE-2	0.2182
ROUGE-L	0.3808
ROUGE-LSum	0.3942
Perplexity	5.7940

Evaluasi terhadap modul *generator* dilakukan untuk menilai kualitas jawaban yang dihasilkan oleh LLM berdasarkan dokumen hasil *retrieval*. Proses ini mencakup pengujian terhadap sejumlah *query* uji yang telah ditentukan, di mana sistem akan menyusun *prompt* yang menggabungkan pertanyaan pengguna dan konteks dokumen, lalu menghasilkan jawaban otomatis.

Berdasarkan hasil evaluasi terhadap modul *generator*, performa sistem dalam menghasilkan jawaban menunjukkan hasil yang cukup baik. Nilai BERT Score sebesar 0.7750 mengindikasikan bahwa secara semantik, jawaban yang dihasilkan memiliki kesamaan makna yang tinggi dengan referensi. Skor ROUGE juga mendukung temuan ini, dengan ROUGE-1 sebesar 0.5032 dan ROUGE-2 sebesar 0.2182, yang menunjukkan bahwa model mampu menangkap sebagian besar *unigram* dan *bigram* penting dari jawaban referensi. Sementara itu, ROUGE-L dan ROUGE-Lsum masing-masing sebesar 0.3808 dan 0.3942 menunjukkan bahwa struktur kalimat model cukup selaras dengan referensi. Di sisi lain, nilai *perplexity* sebesar 5.7940 menunjukkan bahwa model memiliki tingkat ketidakpastian yang cukup rendah dalam menghasilkan token-token jawaban, yang artinya model cukup yakin dengan prediksi kata berikutnya. Secara keseluruhan, metrik-metrik ini mencerminkan bahwa sistem mampu menghasilkan jawaban yang cukup akurat baik secara semantik maupun secara tekstual.

V. KESIMPULAN

Berdasarkan hasil implementasi, evaluasi, dan analisis, dapat disimpulkan bahwa sistem *chatbot e-commerce* berbasis RAG yang dikembangkan mampu menjawab permasalahan utama secara komprehensif. Sistem ini berhasil menghasilkan jawaban yang faktual, relevan, dan kontekstual berkat kombinasi modul *retriever* dan *generator* berbasis LLM. Komponen *retriever* memanfaatkan pencarian semantik berbasis *vector embedding*, sehingga dapat memahami berbagai variasi bentuk pertanyaan pengguna dan menemukan dokumen yang sesuai makna, bukan hanya mencocokkan kata kunci literal. Keberhasilan ini tercermin dari hasil evaluasi menggunakan metrik yang relevan seperti Precision@k, Recall@k, dan *Mean Reciprocal Rank* (MRR) yang menunjukkan kemampuan sistem menemukan dokumen relevan untuk mendukung pembentukan jawaban. Pendekatan ini juga terbukti efektif dalam menjaga konsistensi jawaban berbasis data aktual meskipun *knowledge base* produk diperbarui, tanpa perlu *retraining model* secara penuh, sehingga mendukung efisiensi dan fleksibilitas pengembangan. Hal ini menunjukkan bahwa integrasi *retrieval* dan *generation* dapat memperkuat ketepatan informasi sekaligus mempertahankan relevansi jawaban seiring perubahan data.

Selain itu, modul *generator* berbasis LLM memainkan peran penting dalam menyusun jawaban akhir menjadi kalimat percakapan yang natural, koheren, dan mudah dipahami. *Prompt* yang dirancang sistematis, termasuk instruksi sistem, penyajian konteks terstruktur, dan *user query*, memungkinkan LLM untuk menghasilkan respons yang lebih sesuai konteks dan terdengar lebih komunikatif. Evaluasi menggunakan BERTScore, ROUGE, dan *Perplexity* menunjukkan performa yang cukup baik, mengindikasikan kemiripan semantik dengan jawaban acuan dan kualitas narasi yang responsif. Eksperimen pada variasi *prompt* dan format input juga berhasil meminimalkan jawaban yang kaku atau terlalu umum, sehingga sistem mampu menjawab pertanyaan pengguna dengan lebih responsif dan relevan. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa integrasi *retriever* dan *generator* dalam arsitektur RAG tidak hanya meningkatkan akurasi jawaban, tetapi juga menciptakan pengalaman interaksi yang lebih natural dan informatif bagi pengguna di domain *e-commerce*.

REFERENSI

[1] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, “LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation,” *Natural Language Processing Journal*, vol. 6, p. 100056, Mar. 2024, doi: 10.1016/j.nlp.2024.100056.

[2] S. Angelov and M. Lazarova, “E-commerce distributed chatbot system,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2019. doi: 10.1145/3351556.3351587.

[3] S. Kant Ojha, A. Kumar, T. Bhole, and S. Naaz, “Rule-based A.I. chatbot,” *International Journal of Research & Technology*, vol. 11, Oct. 2024.

[4] J. B. Adelusi, “Hybrid Rule-Based and Machine Learning Chatbots,” *Research Gate*, Aug. 2021, [Online]. Available: https://www.researchgate.net/publication/387669510

[5] M. M. Khan, “Development of An e-commerce Sales Chatbot,” in *HONET 2020 - IEEE 17th International Conference on Smart Communities: Improving Quality of Life using ICT, IoT and AI*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 173–176. doi: 10.1109/HONET50430.2020.9322667.

[6] J. Yun and J. Park, “The Effects of Chatbot Service Recovery With Emotion Words on Customer Satisfaction, Repurchase Intention, and Positive Word-Of-Mouth,” *Front Psychol*, vol. 13, May 2022, doi: 10.3389/fpsyg.2022.922503.

[7] V. Oguntosin and A. Olomo, “Development of an E-Commerce Chatbot for a University Shopping Mall,” *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021, doi: 10.1155/2021/6630326.

[8] H. Naveed *et al.*, “A Comprehensive Overview of Large Language Models,” Oct. 2024, [Online]. Available: http://arxiv.org/abs/2307.06435

[9] D. Venkatachalam, J. Sreerama, S. Analytics, and U. Rajalakshmi Soundarapandiyan, “Large Language Models in Retail: Best Practices for Training, Personalization, and Real-Time Customer Interaction in E-Commerce Platforms,” *Journal of Artificial Intelligence Research and Applications*, vol. 4, Jun. 2024.

[10] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications,” Apr. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15084234.

[11] P. Parmar, “Semantic Search and Question-Answering Systems,” Indian Institute of Science Education and Research Pune, 2023.

[12] M. Douze *et al.*, “The Faiss library,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2401.08281