PROPOSAL TUGAS AKHIR

Perancangan Chatbot Layanan Pelanggan Menggunakan Retrieval Augmented Generation dan Gemini

Nama : Reza Pratama Tandjung
NRP : 221116984
Jurusan / Prodi/ Major : Teknik Informatika / S1 / Intelligent System
Dosen Pembimbing : Esther Irawati, S.Kom, M.Kom

I. Latar Belakang

Di era digital saat ini, teknologi telah mengubah cara perusahaan berinteraksi dengan pelanggan. Salah satu inovasi penting adalah chatbot, program komputer yang dirancang untuk meniru percakapan manusia guna memberikan respons otomatis terhadap pertanyaan atau permintaan pengguna. Chatbot, terutama yang berbasis kecerdasan buatan, mampu meningkatkan efisiensi operasional dan kualitas layanan pelanggan dengan memberikan respons yang cepat dan akurat.

Layanan pelanggan merupakan aspek vital dalam menjaga kepuasan dan loyalitas pelanggan. Dengan meningkatnya ekspektasi terhadap layanan yang cepat dan tepat, perusahaan menghadapi tantangan besar dalam memenuhi kebutuhan tersebut. Chatbot dapat beroperasi 24/7, memberikan solusi instan, mengurangi waktu tunggu, dan menyelesaikan berbagai masalah dengan cepat.

Gemini adalah teknologi terbaru dalam pengembangan chatbot, menawarkan fitur-fitur canggih seperti Natural Language Processing (NLP), machine learning, dan kemampuan integrasi yang luas. Teknologi ini memungkinkan chatbot untuk memahami dan merespons pertanyaan pelanggan dengan lebih baik dan alami, meningkatkan interaksi antara pelanggan dan perusahaan.

Namun, perancangan dan implementasi chatbot layanan pelanggan menggunakan Gemini tidaklah mudah. Dibutuhkan perencanaan yang cermat, pemahaman mendalam mengenai kebutuhan pelanggan, serta pengujian dan evaluasi berkelanjutan untuk memastikan chatbot berfungsi optimal. Aspek keamanan data dan privasi pelanggan juga sangat penting dalam pengembangan chatbot.

Tugas Akhir ini bertujuan untuk merancang chatbot layanan pelanggan menggunakan teknologi Gemini, dengan fokus pada proses desain, implementasi, dan evaluasi performa chatbot dalam meningkatkan kualitas layanan pelanggan di Industri Dengan menggunakan Gemini, diharapkan chatbot yang dikembangkan mampu memberikan solusi efektif dan efisien dalam menangani berbagai kebutuhan dan pertanyaan pelanggan.

Melalui Tugas Akhir ini, diharapkan dapat dihasilkan panduan praktis untuk perancangan dan implementasi chatbot layanan pelanggan berbasis Retrieval Augmented Generation dengan Gemini. Metode ini dipilih karena kemampuannya untuk memastikan chatbot memberikan jawaban yang faktual dengan mengambil informasi langsung dari basis pengetahuan spesifik perusahaan, sehingga mengurangi risiko jawaban yang tidak akurat atau "halusinasi" dari model bahasa. Dengan demikian, penelitian ini dapat memberikan kontribusi nyata dalam pengembangan teknologi layanan pelanggan. Hasil Tugas Akhir ini juga diharapkan dapat membantu perusahaan lain dalam mengadopsi teknologi serupa untuk meningkatkan layanan mereka.

II. Tujuan

Tujuan dari tugas akhir ini akan dijelaskan sebagai berikut:

1. Merancang dan mengembangkan chatbot layanan pelanggan dengan arsitektur Retrieval-Augmented Generation (RAG) menggunakan model bahasa Gemini.
2. Membangun basis pengetahuan (knowledge base) terstruktur dari dokumen perusahaan yang akan digunakan oleh sistem RAG untuk memberikan jawaban yang akurat dan kontekstual.
3. Mengukur efektivitas chatbot dalam mengurangi waktu respons dan meningkatkan kepuasan pelanggan melalui metrik kinerja yang telah ditentukan.

III. Teori Penunjang

Adapun teori penunjang dalam tugas akhir ini adalah sebagai berikut:

A. AI ChatBot

AI chatbot, yang merupakan aplikasi atau antarmuka yang mampu menjalankan percakapan mirip manusia menggunakan natural language processing (NLP) dan machine learning (ML). Berbeda dari chatbot standar, AI chatbot menggunakan large language model (LLM) untuk memahami konteks dan menghasilkan respons yang lebih dinamis terhadap masukan teks.

B. Natural Language Proccessing (NLP)

natural language processing merupakan menggabungkan linguistik komputasional model berbasis aturan bahasa manusia dengan model statistik dan pembelajaran mesin untuk memungkinkan komputer dan perangkat digital untuk mengenali, memahami, dan menghasilkan teks dan ucapan. NLP merupakan cabang dari kecerdasan buatan (AI) yang berperan penting dalam aplikasi dan perangkat yang mampu menerjemahkan teks dari satu bahasa ke bahasa lain, merespons perintah yang diketik atau diucapkan, mengenali atau mengotentikasi pengguna berdasarkan suara, merangkum teks dalam volume besar, menilai niat atau sentimen dari teks atau ucapan, dan menghasilkan teks, grafik, atau konten lainnya secara langsung.

C. Large Language Model (LLM)

Large Language Model (LLM) adalah jenis program kecerdasan buatan yang dapat mengenali dan menghasilkan teks, di antara tugas-tugas lainnya. LLM dilatih menggunakan set data yang besar, sehingga disebut "besar". LLM dibangun di atas pembelajaran mesin: khususnya, jenis jaringan saraf yang disebut model transformer yang memungkinkannya memahami pola, tata bahasa, dan konteks dalam bahasa manusia.

D. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) adalah sebuah arsitektur canggih yang meningkatkan kemampuan LLM dengan menghubungkannya ke sumber pengetahuan eksternal. Pendekatan ini mengatasi kelemahan LLM yang pengetahuannya terbatas pada data pelatihannya dan rentan mengalami "halusinasi" (memberikan informasi yang salah).

Sistem RAG bekerja dalam dua tahap utama:

1. **Retrieval (Pengambilan):** Ketika pengguna mengajukan pertanyaan, sistem terlebih dahulu mencari dan mengambil potongan informasi yang paling relevan dari sebuah basis pengetahuan yang telah disiapkan (misalnya, dokumen internal perusahaan, FAQ, atau manual produk).
2. **Generation (Pembuatan):** Potongan informasi yang relevan tersebut kemudian digabungkan dengan pertanyaan asli pengguna dan diberikan kepada LLM (seperti Gemini) sebagai konteks. LLM kemudian menghasilkan jawaban yang didasarkan pada konteks yang diberikan, sehingga hasilnya jauh lebih akurat dan faktual.

E. Text Embedding

Text embedding adalah proses mengubah teks (kata, kalimat, atau seluruh dokumen) menjadi representasi numerik dalam bentuk vektor. Tujuan utamanya adalah menangkap makna semantik dari teks tersebut. Dalam sistem RAG, pertanyaan pengguna dan potongan dokumen pengetahuan diubah menjadi vektor. Hal ini memungkinkan sistem untuk membandingkan kemiripan makna antara pertanyaan dan dokumen secara matematis untuk menemukan informasi yang paling relevan.

F. Vector Database

Vector database (basis data vektor) adalah sistem basis data yang dirancang khusus untuk menyimpan dan melakukan pencarian pada data vektor berdimensi tinggi, seperti text embeddings. Saat sistem RAG perlu menemukan informasi yang relevan, ia akan mencari vektor pertanyaan pengguna di dalam vector database untuk menemukan vektor dokumen yang paling mirip (paling dekat secara geometris), yang menandakan relevansi konten yang tinggi.

G. Document Chunking

Document chunking adalah proses memecah dokumen besar menjadi potongan-potongan teks yang lebih kecil dan dapat dikelola (*chunks*). Hal ini

penting karena LLM memiliki batasan jumlah teks (panjang konteks) yang dapat diproses sekaligus. Dengan memberikan potongan yang lebih kecil dan relevan, LLM dapat fokus pada informasi yang paling penting untuk menghasilkan jawaban yang akurat.

H. Gemini

Gemini merupakan sistem artificial intelligence (AI) inovatif yang dikembangkan oleh Google dengan memanfaatkan metodologi pelatihan yang berasal dari AlphaGo. Gemini atau yang lebih sering dikenal dengan Google Gemini mampu memahami dan memproses berbagai perintah, termasuk gambar, teks, ucapan, musik, kode komputer, dan banyak lagi. Model AI Gemini disebut sebagai model "Multimodal" yang memungkinkannya melakukan tugas-tugas melebihi kemampuan model bahasa tradisional, seperti menghasilkan gambar dari deskripsi teks atau menerjemahkan antara modalitas yang berbeda.

Dalam arsitektur RAG pada penelitian ini, Gemini berperan sebagai komponen generator yang cerdas, bertugas untuk merangkai jawaban yang koheren dan kontekstual berdasarkan informasi yang diterima dari tahap retrieval.

I. Google Cloud

Google Cloud adalah kumpulan layanan komputasi awan yang ditawarkan oleh Google. Layanan ini menyediakan berbagai alat dan infrastruktur untuk membantu bisnis, pengembang, dan peneliti dalam menjalankan aplikasi dan layanan secara efektif dan efisien di internet. Google Cloud mencakup berbagai produk dan layanan, termasuk penyimpanan data, analisis data, komputasi, dan pembelajaran mesin (machine learning/ML).

J. LangChain

LangChain adalah kerangka kerja *open-source* yang sangat mempermudah pengembangan aplikasi berbasis LLM. LangChain menyediakan serangkaian alat dan API untuk menyederhanakan proses kompleks dalam membangun sistem RAG, seperti memuat dokumen, melakukan chunking, membuat embedding, berinteraksi dengan vector database, dan mengelola seluruh alur kerja dari pertanyaan hingga jawaban.

K. Python

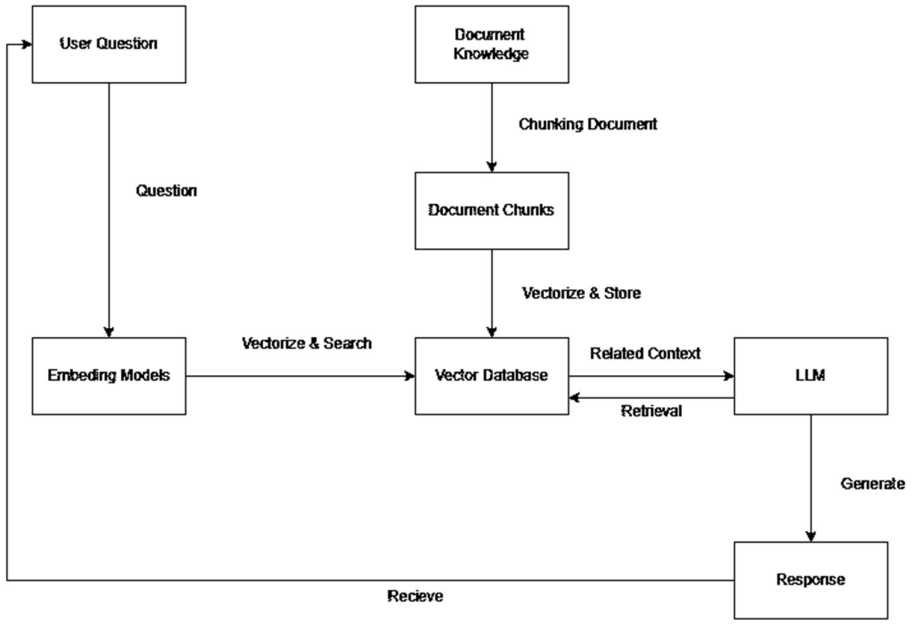
Python adalah bahasa pemrograman tingkat tinggi yang sangat populer di bidang machine learning dan AI. Bahasa ini dipilih karena memiliki ekosistem library yang kaya (seperti LangChain dan Google Cloud), sintaks yang mudah dibaca, dan dukungan komunitas yang besar, menjadikannya pilihan ideal untuk membangun prototipe dan sistem produksi chatbot.

IV. Ruang Lingkup

Bagian ini akan menjelaskan tentang ruang lingkup dari tugas akhir kali ini. Beberapa hal yang akan dikerjakan masing-masing akan dijelaskan sebagai berikut:

A. Arsitektur Sistem

Fokus dari tugas akhir ini adalah Membuat Chatbot Untuk Keperluan Layanan Pelanggan.



Gambar 1
Arsitektur Sistem

Gambar 1 menunjukkan arsitektur Retrieval-Augmented Generation (RAG) sistem dari chatbot pelayanan customer yang akan dibuat pada tugas akhir ini. Sistem ini dirancang agar pengguna (user) dapat menginputkan data perusahaannya sehingga konsumen bisa mendapatkan jawaban yang akurat dan sesuai dengan kebutuhan mereka. Berikut adalah penjelasan blok diagram tersebut:

- User Question: Proses dimulai ketika konsumen memasukkan pertanyaan ke dalam chatbot.
- Embedding Models: Pertanyaan yang diberikan oleh konsumen kemudian diubah menjadi representasi vektor menggunakan model embedding. Model embedding ini bertujuan untuk menangkap konteks dari pertanyaan tersebut.
- Vectorize & Search : Vektor hasil dari embedding model digunakan untuk mencari konteks yang relevan dalam basis data vektor.

- Document Knowledge : Pengetahuan yang diinput oleh user (misalnya, data perusahaan) dibagi menjadi beberapa bagian atau potongan (chunking document) untuk memudahkan pemrosesan lebih lanjut.
- Document Chunks : Setiap potongan dokumen kemudian diubah menjadi representasi vektor dan disimpan dalam basis data vektor.
- Vector Database: Basis data vektor menyimpan representasi vektor dari seluruh potongan dokumen. Ketika pertanyaan dari pengguna datang, basis data ini digunakan untuk menemukan konteks yang relevan dengan melakukan pencarian berdasarkan vektor pertanyaan.
- Retrieval: Konteks yang relevan ditemukan dalam basis data vektor kemudian diambil dan dikirim ke model bahasa besar (LLM).
- LLM: Model bahasa besar menggunakan konteks yang ditemukan untuk menghasilkan respons yang relevan dan informatif.
- Response: Respons yang dihasilkan oleh LLM akan ditampilkan ke pengguna.

Sistem ini memungkinkan perusahaan untuk menginput dan mengelola pengetahuan mereka, yang kemudian dapat digunakan oleh chatbot untuk menjawab pertanyaan konsumen dengan akurat. Dengan menggunakan teknik embedding dan basis data vektor, sistem dapat mencari dan mengambil informasi yang relevan untuk memberikan jawaban yang tepat, sementara model bahasa besar membantu menghasilkan respons yang alami dan kontekstual.

B. Fitur

Fitur fitur yang akan dibuat antara lain :

- Manajemen Basis Pengetahuan (Knowledge Base Management): Ini adalah fitur kunci yang memungkinkan administrator atau pengguna bisnis untuk mengunggah dan mengelola sumber pengetahuan perusahaan (misalnya dalam format PDF, DOCX, atau daftar FAQ). Dokumen-dokumen inilah yang akan menjadi "otak" atau sumber kebenaran bagi chatbot untuk menjawab pertanyaan.
- Jawaban Kontekstual Berbasis Pengetahuan: Chatbot mampu menjawab pertanyaan umum dan spesifik secara otomatis dengan cara mengambil informasi langsung dari basis pengetahuan yang telah diunggah. Ini lebih dari sekadar FAQ statis, karena chatbot dapat mensintesis jawaban dari berbagai sumber dokumen.
- Pemandu Langkah-demi-Langkah: Sistem dapat memberikan panduan terstruktur untuk menyelesaikan masalah umum dengan mengambil prosedur yang relevan dari basis pengetahuan.
- Rekomendasi Produk Cerdas: Chatbot dapat memberikan rekomendasi produk yang disesuaikan dengan kebutuhan pengguna dengan cara mengambil informasi dari katalog produk dan data preferensi yang ada di dalam basis pengetahuan.

- Pengelolaan Konteks Percakapan: Sistem akan menyimpan riwayat percakapan untuk memahami konteks dan memberikan respons yang berkelanjutan dan relevan.
- Alur Percakapan Dinamis: Memanfaatkan kemampuan LLM, chatbot akan memiliki alur percakapan yang fleksibel dan dapat beradaptasi dengan pertanyaan atau perubahan topik dari pengguna secara alami.
- Eskalasi ke Agen Manusia: Jika chatbot tidak dapat menangani permintaan atau jika pengguna secara eksplisit memintanya, sistem akan menyediakan opsi untuk mentransfer percakapan ke agen manusia.
- Pelacakan Kinerja Chatbot: Dasbor akan memantau kinerja chatbot melalui metrik kunci seperti tingkat keberhasilan jawaban, waktu respons, dan skor kepuasan pengguna.
- Pelaporan dan Analisis: Sistem akan menghasilkan laporan dan analisis interaksi pengguna untuk mengidentifikasi pertanyaan yang sering muncul, area di mana chatbot gagal, dan peluang untuk perbaikan.

C. Input dan Output Program

1. Input

Input untuk AI chatbot customer service adalah data atau informasi yang diterima dari pengguna yang berinteraksi dengan chatbot. Input ini bisa dalam bentuk teks atau suara dan mencakup berbagai jenis permintaan atau pertanyaan dari pelanggan.

Contoh Input:

- "Apa status pesanan saya?"
- "Bagaimana cara mengembalikan produk?"
- "Berapa lama pengiriman?"
- "Saya ingin berbicara dengan agen."

2. Proses

Setelah menerima input dari pengguna, chatbot akan memproses informasi melalui serangkaian langkah yang sesuai dengan arsitektur Retrieval-Augmented Generation (RAG):

- Embedding Pertanyaan Pengguna: Pertama, pertanyaan pengguna yang dalam bentuk teks diubah menjadi representasi numerik (vektor) menggunakan embedding model. Proses ini memungkinkan sistem untuk memahami makna semantik dari pertanyaan tersebut.
- Pencarian Informasi Relevan (Retrieval): Vektor pertanyaan kemudian digunakan untuk melakukan pencarian di dalam *vector database*. Sistem akan mencari dan mengambil beberapa potongan dokumen (*chunks*) dari basis pengetahuan yang memiliki kemiripan makna tertinggi dengan pertanyaan pengguna.
- Augmentasi Konteks untuk LLM: Potongan-potongan dokumen yang relevan tersebut digabungkan dengan pertanyaan asli pengguna untuk membentuk sebuah prompt yang kaya akan konteks.

- Generasi Respons oleh LLM: Prompt yang sudah diperkaya dengan konteks ini kemudian dikirim ke Large Language Model (Gemini). Gemini akan menggunakan informasi yang diberikan untuk menghasilkan jawaban yang akurat, relevan, dan faktual dalam bahasa yang alami.
- Manajemen Dialog: Selama interaksi, sistem akan mengelola riwayat percakapan untuk menjaga konteks, memungkinkan percakapan yang berkelanjutan dan koheren.

3. Output

Output dari AI chatbot adalah respons atau tindakan yang diberikan chatbot sebagai tanggapan terhadap input dari pengguna. Respons ini dirancang untuk menyelesaikan pertanyaan atau permintaan pengguna dengan cara yang paling efisien dan memuaskan.

Contoh Output:

- Untuk input "Apa status pesanan saya?", outputnya bisa berupa: "Pesanan Anda dengan nomor ID 12345 sedang dalam proses pengiriman dan diperkirakan akan tiba dalam 3 hari kerja."
- Untuk input "Bagaimana cara mengembalikan produk?", outputnya bisa berupa: "Untuk mengembalikan produk, Anda dapat mengunjungi halaman pengembalian di situs kami dan mengikuti instruksi yang diberikan. Apakah Anda memerlukan bantuan lebih lanjut?"
- Untuk input "Berapa lama pengiriman?", outputnya bisa berupa: "Pengiriman standar memakan waktu antara 5-7 hari kerja. Anda juga dapat memilih pengiriman ekspres yang memakan waktu 2-3 hari kerja."
- Untuk input "Saya ingin berbicara dengan agen.", outputnya bisa berupa: "Tolong tunggu sebentar, saya akan menghubungkan Anda dengan agen manusia. Mohon bersabar."

4. Contoh Interaksi

Pengguna: "Apa status pesanan saya?"

Chatbot: Selamat datang di layanan pelanggan! Tolong berikan nomor pesanan Anda.

Pengguna: "12345"

Chatbot: Pesanan Anda dengan nomor ID 12345 sedang dalam proses pengiriman dan diperkirakan akan tiba dalam 3 hari kerja. Ada yang bisa saya bantu lagi?

Pengguna: "Tidak, terima kasih."

Chatbot: Terima kasih telah menghubungi layanan pelanggan. Selamat hari!

Dalam contoh interaksi ini, input awal dari pengguna memicu chatbot untuk meminta informasi tambahan (nomor pesanan). Setelah menerima nomor pesanan, chatbot menjalankan alur kerja RAG: sistem akan mencari informasi pesanan tersebut dari basis data pengetahuan, lalu menggunakan

data yang ditemukan sebagai konteks untuk menghasilkan respons yang relevan mengenai status pengiriman.

D. Uji Coba

Setelah pengembangan program chatbot layanan pelanggan menggunakan Gemini selesai, perlu dilakukan uji coba komprehensif untuk memastikan fungsionalitas, keandalan, dan kepuasan pengguna. Target uji coba utama adalah untuk mengevaluasi kemampuan chatbot dalam memahami dan merespons berbagai pertanyaan dan permintaan dari pengguna secara tepat dan efisien. Pengujian ini akan mencakup beberapa skenario yang mencerminkan situasi nyata yang mungkin dihadapi chatbot saat digunakan oleh pelanggan.

1. Target Uji Coba:
   - Akurasi Respons: Mengukur seberapa tepat chatbot dalam memberikan jawaban terhadap pertanyaan pengguna.
   - Kecepatan Respons: Menilai waktu yang dibutuhkan chatbot untuk merespons pertanyaan atau permintaan pengguna.
   - Kepuasan Pengguna: Mengumpulkan umpan balik dari pengguna untuk menilai tingkat kepuasan mereka terhadap interaksi dengan chatbot.
   - Stabilitas Sistem: Menilai kinerja chatbot di bawah berbagai kondisi beban kerja untuk memastikan stabilitas dan keandalan.
   - Pemahaman Konteks: Menguji kemampuan chatbot untuk memahami dan mempertahankan konteks percakapan dalam interaksi yang berkelanjutan.
   - Akurasi Retrieval: Mengukur seberapa akurat sistem dalam menemukan dan mengambil potongan dokumen (*chunks*) yang paling relevan dari basis pengetahuan. Target ini penting untuk memastikan bahwa konteks yang diberikan kepada LLM sudah tepat.
   - Kualitas dan Kepatuhan Generasi (Faithfulness): Menilai seberapa baik LLM (Gemini) dalam menghasilkan jawaban yang didasarkan hanya pada konteks yang diambil. Ini untuk mengukur apakah chatbot tetap faktual sesuai sumber dan tidak "berhalusinasi" atau menambahkan informasi dari luar.
2. Skenario Uji Coba:
   - Skenario Pertanyaan Umum: Menguji respons chatbot terhadap pertanyaan yang sering diajukan, seperti "Bagaimana cara mengembalikan produk?" atau "Berapa lama waktu pengiriman?" untuk memastikan chatbot dapat memberikan jawaban yang benar dan konsisten.
   - Skenario Masalah Teknis: Menguji bagaimana chatbot menangani masalah teknis yang dilaporkan oleh pengguna, seperti "Aplikasi saya tidak bisa dibuka" atau "Saya tidak bisa masuk ke akun saya."
   - Skenario Pemesanan dan Status Pesanan: Menguji kemampuan chatbot untuk menangani pertanyaan terkait pemesanan dan status pesanan, seperti "Apa status pesanan saya?" atau "Bisakah saya membatalkan pesanan saya?"

- Skenario Interaksi Kompleks: Menguji kemampuan chatbot untuk menangani percakapan yang lebih kompleks dan berkelanjutan, di mana pengguna mengajukan beberapa pertanyaan dalam satu sesi, seperti "Bagaimana cara mengembalikan produk?" diikuti dengan "Apa kebijakan pengembalian dana?"
- Skenario Pengalihan ke Agen Manusia: Menguji fitur pengalihan ke agen manusia saat chatbot tidak dapat menyelesaikan permintaan pengguna atau saat pengguna secara khusus meminta untuk berbicara dengan agen, seperti "Saya ingin berbicara dengan agen" atau "Saya tidak puas dengan jawaban ini."
- Skenario Pertanyaan di Luar Pengetahuan: Menguji respons chatbot ketika diberikan pertanyaan yang jawabannya tidak ada di dalam basis pengetahuan. Respons yang ideal adalah chatbot menyatakan tidak tahu atau tidak dapat menemukan informasi, bukan mencoba menjawab dengan salah.

Uji coba akan melibatkan pengguna internal dan eksternal untuk memastikan bahwa chatbot dapat beroperasi secara optimal dalam berbagai situasi nyata. Hasil uji coba ini akan dianalisis untuk mengidentifikasi area yang perlu perbaikan dan untuk memastikan bahwa chatbot dapat memberikan layanan pelanggan yang efektif dan memuaskan.

V. Tahapan Penyelesaian Tugas Akhir

1. Perencanaan dan Tugas Akhir Awal: Pada tahap awal, dilakukan kajian literatur untuk memahami konsep dasar dan teknologi yang akan digunakan, seperti RAG, LLM, dan vector database. Selain itu, dilakukan analisis kebutuhan fungsional dan non-fungsional dari chatbot layanan pelanggan. Pengumpulan data primer melalui wawancara atau survei dapat dilakukan untuk mendapatkan wawasan mendalam tentang kebutuhan dan harapan pengguna terhadap chatbot.
2. Pengumpulan dan Penyiapan Basis Pengetahuan: Tahap ini berfokus pada pengumpulan sumber informasi yang akan menjadi "otak" bagi chatbot. Data yang dikumpulkan bukanlah dataset untuk melatih model, melainkan dokumen-dokumen sumber seperti:
    - Manual produk dalam format PDF atau DOCX.
    - Daftar Pertanyaan yang Sering Diajukan (FAQ).
    - Dokumen kebijakan perusahaan.
    - Artikel atau panduan internal lainnya. Data ini kemudian dibersihkan dan distrukturkan agar siap diproses oleh sistem RAG.
3. Implementasi Pipeline RAG dan Pengembangan Chatbot : Tahap ini adalah inti dari pengembangan teknis, yang meliputi:
    - Pemrosesan Dokumen: Mengimplementasikan proses untuk memuat dokumen dari basis pengetahuan dan memecahnya menjadi potongan-potongan yang lebih kecil (document chunking).

- Pembuatan Vector Store: Memilih embedding model yang sesuai untuk mengubah setiap potongan dokumen menjadi vektor, kemudian menyimpannya ke dalam vector database.
- Pembangunan Alur RAG: Mengembangkan logika utama sistem menggunakan kerangka kerja seperti LangChain. Ini mencakup alur yang menerima pertanyaan pengguna, membuat embedding dari pertanyaan tersebut, mengambil dokumen relevan dari vector database, dan menggabungkannya menjadi prompt untuk LLM.
- Integrasi dengan Gemini: Menghubungkan alur RAG dengan model Gemini, yang akan bertindak sebagai generator untuk menghasilkan jawaban akhir berdasarkan konteks yang diberikan.
- Pengembangan Antarmuka Pengguna (Frontend): Membangun antarmuka chatbot tempat pengguna dapat berinteraksi dengan sistem.

4. Pengujian Sistem: Setelah pengembangan, dilakukan pengujian menyeluruh untuk memastikan bahwa chatbot berfungsi sesuai harapan. Pengujian ini meliputi:
   - Uji Fungsional: Menguji setiap fitur dan fungsi chatbot untuk memastikan semuanya bekerja dengan baik.
   - Uji Kinerja: Mengukur kecepatan dan responsivitas chatbot di bawah berbagai kondisi beban.
   - Uji Pengguna: Melibatkan pengguna nyata untuk berinteraksi dengan chatbot dan memberikan umpan balik mengenai pengalaman mereka. Ini mencakup uji coba skenario yang telah ditentukan sebelumnya untuk mengevaluasi keakuratan dan relevansi respons chatbot.
   - Pengujian spesifik RAG untuk mengukur akurasi retrieval dan kepatuhan jawaban (faithfulness).
5. Evaluasi dan Validasi: Tahap akhir melibatkan evaluasi hasil pengujian untuk mengidentifikasi area yang perlu perbaikan. Umpan balik dari pengguna digunakan untuk melakukan iterasi pada desain dan fungsionalitas chatbot. Validasi dilakukan dengan membandingkan kinerja chatbot dengan standar industri dan kebutuhan pengguna. Metode evaluasi mencakup analisis kuantitatif (seperti tingkat keberhasilan respons dan waktu respons) serta analisis kualitatif (seperti kepuasan pengguna).
6. Dokumentasi dan Pelaporan: Hasil dari seluruh tahapan Tugas Akhir didokumentasikan secara rinci. Laporan Tugas Akhir mencakup metodologi yang digunakan, hasil pengujian, analisis data, dan rekomendasi untuk pengembangan lebih lanjut. Dokumentasi ini penting untuk memastikan bahwa proses dan hasil Tugas Akhir dapat dipertanggungjawabkan dan diimplementasikan di masa depan.
7. Hosting Ke Google Cloud: Tahap akhir melibatkan hosting chatbot ke Google Cloud untuk memastikan ketersediaan dan skalabilitasnya dalam lingkungan produksi. Dengan hosting di Google Cloud, chatbot dapat

dioperasikan dengan efisien, mendukung akses yang aman dan cepat oleh pengguna. Hosting ini juga memungkinkan pemantauan dan pemeliharaan yang lebih baik serta dukungan untuk skala yang lebih besar sesuai kebutuhan pengguna.

VI. Daftar Pustaka

1. Vaswani, A., et al, 2017. *Attention Is All You Need*. Alamat Web: https://arxiv.org/abs/1706.03762.
2. Lewis, P., et al, 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Alamat Web: https://arxiv.org/abs/2005.11401.
3. Gemini Team, Google, 2023. *Gemini: A Family of Highly Capable Multimodal Models. Alamat Web: https://arxiv.org/abs/2312.11805.*
4. Es, S., et al, 2023. *RAGAs: Automated Evaluation of Retrieval Augmented Generation*. Alamat Web: https://arxiv.org/abs/2309.15217.
5. Kshitiz Rimal, 2024. *MultiModal RAG with Gemini Pro and Langchain*. Alamat Web: https://medium.com/next-ai/multimodal-rag-with-gemini-pro-and-langchain-e4f74170420a.
6. Bill Huang, 2024. *Chatbot with LLM and RAG in action*. Alamat Web: https://medium.com/@yingbiao/chatbot-with-llm-and-rag-in-action-575382df4323.
7. Florian June, 2024. *A brief introduction to retrieval augmented generation(RAG)*. Alamat Web: https://ai.plainenglish.io/a-brief-introduction-to-retrieval-augmented-generation-rag-b7eb70982891.
8. Google, 2024. *Google AI for Developers - Gemini API Documentation*. Alamat Web: https://ai.google.dev/docs.
9. LangChain, 2024. *LangChain Python Documentation*. Alamat Web: https://python.langchain.com/.