BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Chatbot merupakan aplikasi yang berfungsi sistem percakapan otomatis yang beroperasi melalui interaksi kecerdasan buatan dengan pengguna melalui Natural Language Processing (NLP). Penggunaan chatbot semakin meluas dalam berbagai sektor global seperti kesehatan, keuangan, pendidikan, pertanian, dan industri lainnya [1].

Perkembangan zaman telah mempercepat adopsi chatbot dalam berbagai industri. Salah satu kontributor utama adalah kemajuan dalam bidang *Artificial Intelligence*, terutama dengan kemunculan teknologi seperti *Chat GPT* dari *OpenAI*. *Chat GPT* merupakan *Large Language Model (LLM)* yang dapat diintegrasikan dengan sistem *chatbot* untuk menghasilkan jawaban yang lebih relevan dan menyesuaikan dengan pertanyaan yang diberikan oleh pengguna [2]. Dengan melakukan integrasi antara chatbot dengan teknologi dari Chat GPT, hal tersebut memungkinkan untuk pengembangan sistem yang lebih inovatif. Dengan demikian, perkembangan zaman telah mempercepat integrasi chatbot berbasis Chat GPT dalam berbagai sektor, memungkinkan mereka untuk mengembangkan sistem percakapan otomatis yang lebih efektif dan canggih .

Chat GPT yang dikembangkan oleh OpenAI merupakan sebuah Large Language Model (LLM) yang memberikan terobosan baru dalam dunia *Generative AI* berbasis teks dimana teknologi tersebut memiliki kemampuan untuk menghasilkan jawaban yang relevan berdasarkan pertanyaan yang diberikan oleh manusia. Meskipun *LLM* menawarkan banyak keuntungan, teknologi tersebut juga memiliki kelemahan dimana salah satunya adalah teknologi tersebut memiliki kecenderungan untuk ”berhalusinasi” atau menghasilkan informasi yang tidak akurat atau sepenuhnya dibuat-buat namun seolah-olah mereka yakin akan jawaban tersebut [3].

Halusinasi dalam bidang *Artificial Intelligence* merujuk pada situasi di mana model AI menghasilkan informasi atau jawaban yang tampak meyakinkan namun tidak akurat atau tidak sesuai dengan fakta yang sebenarnya [4]. Oleh karena itu, sangat penting bagi pengguna untuk melakukan verifikasi terlebih dahulu informasi yang diberikan oleh AI dengan membandingkannya dengan sumber data yang

valid, relevan, dan terkini informasinya guna menghindari potensi dampak negatif dari halusinasi ini mengingat keakuratan informasi dan relevansi jawaban yang diberikan oleh AI sangat penting [5].

Metode Retrieval Augmented Generation (RAG) mengusulkan solusi untuk masalah ini dengan menggabungkan keunggulan dari *LLM* dengan mekanisme pengambilan data berbasis *text query* milik *vector database* dimana sebelum AI menjawab pertanyaan yang diberikan oleh pengguna, sistem akan mencari informasi yang relevan terlebih dahulu dalam *database* kemudian menambahkan informasi tersebut kedalam konteks pertanyaan pengguna supaya jawaban yang dihasilkan oleh AI selalu relevan berdasarkan data yang ada di *database* [6].

Pemilihan metode RAG dalam penelitian ini didorong oleh keunggulan metode tersebut dalam menggabungkan kemampuan *retrieval*, yaitu mencari informasi yang relevan melalui *vector database* kemudian akan dilakukan *generation* dengan bantuan *LLM (Large Language Modelling)* untuk menghasilkan jawaban yang kontekstual dan informatif [6]. *Vector Database* dapat dimanfaatkan untuk menyediakan informasi yang terkini karena sumber jawaban yang dihasilkan oleh *LLM* berasal dari *pre-training data* dan juga tambahan konteks yang relevan dari hasil *query Vector Database* yang dapat menghilangkan sifat halusinasi dari *AI* untuk menghasilkan suatu jawaban. Selain itu keunggulan teknik *Retrieval Augmented Generation* juga memberikan fleksibilitas terhadap pengetahuan dari AI yang dapat dikontrol pengetahuannya tanpa perlu melakukan training ulang ketika ingin menambah pengetahuan baru [7].

Penelitian terdahulu telah mengembangkan sebuah chatbot *Artificial Intelligence* menggunakan teknik *Retrieval Augmented Generation (RAG)* dimana metodologi penelitian sebelumnya dilakukan dengan memanfaatkan database berbasis graph, khususnya menggunakan *NEO4J* sebagai infrastruktur dasar dimana dalam penelitian tersebut menghasilkan sebuah sistem *chatbot* menggunakan *AI* dengan teknik *RAG* dan mendapatkan akurasi sebesar 90% [8]. Namun, dalam penelitian ini, pendekatan yang berbeda diambil dengan menggunakan vector database untuk implementasi teknik RAG. Hal ini memunculkan perbedaan signifikan dalam pendekatan dan infrastruktur yang digunakan.

Penelitian lainnya juga mengamati tentang analisa vector database dimana database jenis ini secara teori memungkinkan untuk menghilangkan sifat halusinasi dari *AI* dan dapat diimplementasikan melalui teknik *Retrieval Augmented Generation (RAG)* dimana hal ini memungkinkan sistem untuk mengambil dan memanipulasi informasi yang diperlukan [9]. Dengan menggunakan pendekatan

ini, penelitian sebelumnya telah menunjukkan potensi besar dalam membangun chatbot yang lebih responsif dan efektif dengan memanfaatkan fitur-fitur khusus dari vector database, seperti kemampuan untuk menghitung kemiripan antara vektor. penelitian ini bertujuan untuk mengeksplorasi kemampuan, kinerja, dan akurasi teknik *RAG* ketika diterapkan dengan menggunakan vector database [6]. Dalam teknik *RAG*, pengetahuan tidak perlu dimasukkan kedalam parameter model *LLM* melainkan dapat diperoleh hanya dengan pengambilan data secara eksplisit dari luar dengan gaya *plug-and-play* sehingga menciptakan skalabilitas yang besar dalam pengembangannya [10]. Dengan demikian, penelitian ini diharapkan dapat memberikan wawasan yang lebih mendalam mengenai efektivitas dan efisiensi penggunaan database berbasis vector dalam bidang *AI*.

Dalam sebuah penelitian, mengidentifikasi research gap merupakan langkah krusial untuk menentukan arah penelitian yang relevan dan signifikan. Research gap adalah celah atau kekurangan dalam pengetahuan yang ada dalam bidang keilmuan tertentu yang belum terjawab oleh penelitian-penelitian sebelumnya. Dengan memahami gap ini, peneliti dapat merumuskan pertanyaan penelitian yang lebih tepat dan berkontribusi secara signifikan pada pengembangan ilmu pengetahuan. Penelitian-penelitian terdahulu akan disajikan dalam bentuk tabel untuk memudahkan analisis dan perbandingan yang bisa dilihat pada daftar berikut:

1. A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM
    - Distance Metric: Euclidean Distance
    - Chunk Size: 2000 Token
    - Retrieved chunk per request: 5 chunk
    - LLM Model: GPT 3.5 Turbo
    - Embedding Model: tidak disebutkan dalam jurnal
    - Metode Evaluasi: tidak disebutkan dalam jurnal
    - Score: tidak disebutkan dalam jurnal
2. PENERAPAN RETRIEVAL AUGEMENTED GENERATION MENGGUNAKAN LANGCHAIN DALAM PENGEMBANGAN SISTEM TANYA JAWAB HADIS BERBASIS WEB
    - Distance Metric: Euclidean Distance

- Chunk Size: Tidak disebutkan dalam jurnal
- Retrieved chunk per request: Tidak disebutkan dalam jurnal
- LLM Model: GPT-4-1106-preview
- Embedding Model: Text-Embedding-3-Large
- Metode Evaluasi: BERTScore
- Score: F1 Score (0.7962), Akurasi (89.4%)

3. Uji performa chatbot dengan retrieval augmented generation dan model gpt-4 untuk domain taharah berdasarkan empat imam mazhab fikih (studi kasus kitab rahmah al ummah ikhtilaf al a'immah)

    - Distance Metric: Jaccard & Cosine Similarity
    - Chunk Size: Tidak dijelaskan dalam jurnal
    - Retrieved chunk per request: Tidak disebutkan dalam jurnal
    - LLM Model: GPT-4
    - Embedding Model: Tidak disebutkan dalam jurnal
    - Metode Evaluasi: Confusion Matrix
    - Score: Akurasi (90%), Presisi (86.67%), Recall (100%), F1-Score (92.86%)

4. GastroBot: a Chinese gastrointestinal disease chatbot based on the retrievalaugmented generation

    - Distance Metric: Tidak disebutkan dalam jurnal
    - Chunk Size: 512 Character
    - Retrieved chunk per request: 3 chunks
    - LLM Model: Fine Tuned gte-base-zh
    - Embedding Model: Tidak disebutkan dalam jurnal
    - Metode Evaluasi: RAGAS
    - Score: Recall (95%), Faithfulness (93.73%), Relevance (92.28%)

5. Implementasi Teknik Retrieval Augmented Generation untuk Menghilangkan Halusinasi pada Large Language Model Berbasis Vector Database

- Distance Metric: Manhattan
- Chunk Size: 7000 Token
- Retrieved chunk per request: 2 chunks
- LLM Model: GPT-3.5-Turbo
- Embedding Model: text-embedding-ada-002
- Metode Evaluasi: DeepEval (G-Eval)
- Score: Correctness (84.333%)

6. Implementasi Teknik Retrieval Augmented Generation untuk Menghilangkan Halusinasi pada Large Language Model Berbasis Vector Database

    - Distance Metric: Cosine Similarity
    - Chunk Size: 7000 Token
    - Retrieved chunk per request: 2 chunks
    - LLM Model: GPT-3.5-Turbo
    - Embedding Model: text-embedding-ada-002
    - Metode Evaluasi: DeepEval (G-Eval)
    - Score: Correctness (83.333%)

7. Implementasi Teknik Retrieval Augmented Generation untuk Menghilangkan Halusinasi pada Large Language Model Berbasis Vector Database

    - Distance Metric: Euclidean Distance
    - Chunk Size: 7000 Token
    - Retrieved chunk per request: 2 chunks
    - LLM Model: GPT-3.5-Turbo
    - Embedding Model: text-embedding-ada-002
    - Metode Evaluasi: DeepEval (G-Eval)
    - Score: Correctness (86.667%)

8. Implementasi Teknik Retrieval Augmented Generation untuk Menghilangkan Halusinasi pada Large Language Model Berbasis Vector Database

- Distance Metric: Euclidean Distance
- Chunk Size: 2000 Token
- Retrieved chunk per request: 7 chunks
- LLM Model: GPT-3.5-Turbo
- Embedding Model: text-embedding-ada-002
- Metode Evaluasi: DeepEval (G-Eval)
- Score: Correctness (74%)

9. Implementasi Teknik Retrieval Augmented Generation untuk Menghilangkan Halusinasi pada Large Language Model Berbasis Vector Database

    - Distance Metric: Euclidean Distance
    - Chunk Size: 7000 Token
    - Retrieved chunk per request: 2 chunks
    - LLM Model: GPT-3.5-Turbo
    - Embedding Model: text-embedding-3-small
    - Metode Evaluasi: DeepEval (G-Eval)
    - Score: Correctness (78%)

Berdasarkan studi literatur yang telah dilakukan, dalam penelitian ini terdapat perbandingan *score correctness G-Eval* untuk *hyper parameter tuning* terhadap *distance metric, chunk size, retrieved chunk,* serta *embedding modelnya*. Pada penlitian ini, distance metric yang dilakukan pengetesan yaitu *manhattan*, *euclidean distance*, dan juga *cosine similarity* yang akan digunakan untuk menghitung jarak antar vector setiap dokumen yang ada dalam *vector database*, selain itu *chunk size* yang digunakan dalam penelitian ini yaitu antara 2000 dan 7000 token untuk setiap chunk serta embedding yang digunakan yaitu text-embedding-ada-002 dan text-embedding-3-small

Berdasarkan latar belakang tersebut, dalam penelitian ini akan dikembangkan sebuah sistem chat LLM dengan menggunakan metode Retrieval Augmented Generation (RAG) dan vector database dimana hal tersebut dapat mengurangi kecenderungan model untuk berhalusinasi dengan menyediakan konteks yang lebih spesifik dan informasi yang telah diverifikasi, sehingga meningkatkan keakuratan jawaban yang dihasilkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disajikan, didapatkan rumusan masalah sebagai berikut

1. Bagaimana implementasi metode *Retrieval Augmented Generation* untuk menghilangkan halusinasi pada Large Language Model berbasis Vector Databse ?
2. Sejauh mana keakuratan jawaban yang dihasilkan dengan metode Retrieval Augmented Generation (RAG) untuk menghilangkan halusinasi pada Large Language Model berbasis vector database ?

1.3 Batasan Masalah

1. Penelitian ini akan difokuskan pada penerapan aplikasi chat dengan kecerdasan buatan menggunakan *LLM* milik *OpenAI* yaitu *GPT* dan menggabungkan hal tersebut dengan *metode Retrieval Augmented Generation (RAG)* untuk meningkatkan akurasi dari hasil jawaban yang diberikan oleh AI.
2. Database yang akan digunakan dalam pembuatan sistem ini adalah *Vector Database open source* milik Weaviate dan *API LLM* berbayar dengan *GPT* milik *OpenAI*
3. Sistem operasi yang digunakan dalam penelitian ini adalah *Linux* dimana saat ini *Vector Database* milik *Weaviate* hanya *support* untuk *Operating System* berbasis *Linux* saja.
4. Pengujian aplikasi akan memanfaatkan dataset buku PDF Algoritma dan Struktur Data dari Universitas Birmingham, UK untuk menguji keakuratan dari sistem yang dibuat.
5. Peningkatan yang dimaksud dalam penelitian ini adalah peningkatan akurasi antara sistem Chat-GPT sebelum dilakukan implementasi teknik RAG dengan sistem Chat-GPT setelah dilakukan implementasi teknik RAG.
6. Keakuratan yang diukur dalam penelitian ini adalah keakuratan akurasi, presisi, recall, dan F1-Score dimana sistem akan diberikan sebanyak 40 pertanyaan dan akan dibagi kategori nya kedalam beberapa kateogri yaitu:

TP (True Positive), TN (True Negative), FP (False Positive), dan juga FN (False Negative).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk mengimplementasikan metode *Retrieval Augmented Generation (RAG)* untuk menghilangkan halusinasi pada *Large Language Model* berbasis *Vector Database*.
2. Untuk mengukur tingkat akurasi jawaban yang dihasilkan oleh aplikasi tanya jawab AI *Large Language Model* dengan metode *Retrieval Augmented Generation* berbasis *Vector Database*

1.5 Manfaat Penelitian

1. Manfaat bagi Peneliti

    - Pemahaman Mendalam tentang AI dan RAG

      Peneliti akan mendapatkan pemahaman yang mendalam tentang cara kerja kecerdasan buatan, khususnya metode Retrieval Augmented Generation. Ini termasuk pemahaman tentang bagaimana AI dapat digunakan untuk menggabungkan informasi dari database vektor dengan pemrosesan bahasa alami untuk menghasilkan jawaban yang informatif dan relevan.
    - Kontribusi pada Bidang Ilmu Pengetahuan

      Penelitian ini akan memberikan kontribusi penting pada bidang kecerdasan buatan dan pengelolaan sumber daya manusia. Temuan dan metodologi yang dikembangkan dapat menjadi referensi untuk penelitian masa depan dan aplikasi praktis lainnya.

2. Manfaat bagi Pengguna

    - Efisiensi Pemahaman Bisnis

      Aplikasi ini memungkinkan karyawan baru untuk dengan cepat memahami proses bisnis perusahaan tanpa perlu menghabiskan waktu

berjam-jam membaca dokumen. Ini meningkatkan efisiensi dan mempersingkat waktu pembelajaran.

- Peningkatan Produktivitas

  Dengan meminimalkan waktu yang dibutuhkan untuk mencari dan memahami informasi, karyawan dapat lebih fokus pada tugas-tugas penting lainnya, yang pada gilirannya meningkatkan produktivitas kerja.

1.6 Sistematika Penulisan

- Bab 1 PENDAHULUAN

  Bab ini mencakup latar belakang penelitian, identifikasi permasalahan penelitian, pembatasan lingkup investigasi, artikulasi tujuan studi, penjelasan signifikansi penelitian, dan penguraian struktur tesis.

- Bab 2 LANDASAN TEORI

  Bab ini dikhususkan untuk tinjauan literatur relevan yang berfungsi sebagai referensi dasar dan acuan untuk mengembangkan penelitian yang disajikan dalam disertasi ini.

- Bab 3 METODOLOGI PENELITIAN

  Bab metodologi mendetailkan metode penelitian yang digunakan selama studi, termasuk tinjauan literatur, teknik pengumpulan data, desain algoritma, implementasi algoritma, pengembangan situs web, prosedur pengujian dan evaluasi, dokumentasi, dan spesifikasi sistem.

- Bab 4 HASIL DAN DISKUSI

  Bab ini menyajikan temuan penelitian, khususnya algoritma dan situs web yang dikembangkan, diikuti oleh analisis hasil pengujian dan evaluasi dari temuan tersebut.

- Bab 5 KESIMPULAN DAN SARAN

  Bab penutup menyimpulkan temuan dan menawarkan rekomendasi untuk penelitian selanjutnya mengenai topik tersebut, menyarankan jalur untuk penelitian dan pengembangan lebih lanjut.