BAB I
PENDAHULUAN

1.1 Latar Belakang Penelitian

Di era digital saat ini, perkembangan teknologi *Artificial Intelligence (AI)* telah menjadi salah satu pendorong transformasi di berbagai sektor kehidupan, termasuk dalam bidang komunikasi manusia dengan sistem informasi. Salah satu bentuk implementasi AI yang digunakan adalah *chatbot*. *Chatbot* digunakan secara luas di berbagai platform, seperti situs web layanan pelanggan, aplikasi mobile, *e-learning*, dan layanan konsultasi berbasis teks. Awalnya *chatbot* dibangun menggunakan pendekatan *rule-based* yang hanya mampu mengembalikan respon terbatas berdasarkan skenario yang ditentukan secara eksplisit oleh pengembang. Meskipun pendekatan ini sederhana, efektivitasnya rendah karena tidak mampu beradaptasi terhadap variasi input pengguna yang kompleks dan tidak mendukung pemahaman semantik.

Seiring berkembangnya teknologi *Natural Language Processing (NLP)*, muncul sebuah perkembangan baru yaitu *Large Language Models (LLM)* seperti *GPT-3*, *BERT*, dan *LLaMA* yang mendorong evolusi kemampuan chatbot yang lebih kompleks. *LLM* memungkinkan chatbot memahami konteks percakapan dan memberikan respon yang lebih koheren, natural, dan bervariasi. *LLM* dilatih menggunakan dataset besar dari internet dan mampu mempelajari pola linguistik serta informasi faktual dari data tersebut [1]. Hal ini menjadikan chatbot berbasis *LLM* sangat menjanjikan dalam membantu tugas-tugas berbasis teks seperti menjawab pertanyaan, menyarankan solusi, atau menjelaskan konsep tertentu dengan bahasa yang menyerupai manusia.

LLM memiliki keterbatasan fundamental, yaitu bersifat *closed-book,* artinya hanya mengandalkan pengetahuan yang dimilikinya saat pelatihan. Hal ini menyebabkan model tidak memiliki akses terhadap informasi baru atau spesifik yang tidak termasuk dalam data pelatihannya. Akibatnya, model sering mengalami *hallucination* atau memberikan jawaban yang terdengar meyakinkan tetapi sebenarnya salah atau tidak didukung oleh fakta [2]. *Hallucination* pada LLM

adalah fenomena dimana model menghasilkan informasi yang tidak akurat, tidak dapat diverifikasi, atau bahkan sepenuhnya salah, meskipun disajikan dengan bahasa yang tampak meyakinkan dan natural [3]. Dalam konteks penerapan praktis, seperti *chatbot* untuk layanan edukasi atau konsultasi berbasis dokumen resmi, fenomena ini bisa menimbulkan permasalahan serius akibat potensi informasi yang tidak akurat dan membingungkan pengguna. Selain itu, *LLM* yang dilatih pada korpus umum seringkali tidak memiliki cukup informasi domain-spesifik, misalnya tentang regulasi lokal, terminologi teknis, atau dokumen formal tertentu.

Chatbot berbasis LLM masih menghadapi tantangan serius terkait kemampuan memberikan jawaban yang faktual dan dapat diverifikasi. Pendekatan *grounding* seperti yang dilakukan *WikiChat* dengan menggabungkan penyajian faktual dari Wikipedia berhasil meningkatkan faktualitas hingga 97,9%, dibandingkan model LLM murni yang hanya mencapai sekitar 60%-70% akurasi faktual [4]. *Hallucination* tetap jadi tantangan utama dalam LLM. Dalam benchmark *HaluEval* 2.0, ditemukan bahwa LLM masih sering menghasilkan konten faktual salah [5]. Metode *Chain-of-Verification (CoVe)* juga menunjukkan efektivitas dalam menurunkan *hallucination* dengan melakukan verifikasi fakta sebelum menghasilkan jawaban akhir [6]. Hal ini menunjukkan bahwa meskipun LLM mampu menghasilkan teks yang mengalir, tanpa sumber eksternal, chatbot rentan memberikan jawaban yang terdengar meyakinkan namun tidak sesuai fakta.

Penelitian ini menerapkan pendekatan *Retrieval-Augmented Generation (RAG)*. *RAG* menggabungkan kemampuan generatif dari *LLM* dengan teknik pencarian informasi (*retrieval*) dari sumber eksternal. Model ini bekerja dengan dua tahap: pertama, mengambil dokumen atau informasi yang relevan dari basis data menggunakan *vector search* atau *dense retrieval*, lalu kedua, menggabungkan hasil pencarian tersebut kedalam proses generatif untuk menghasilkan respons berbasis konteks yang akurat [7]. Dengan pendekatan ini, *chatbot* tidak hanya mengandalkan memorinya sendiri, tetapi juga dapat "membaca ulang" dokumen dan merujuk langsung pada sumber yang relevan untuk menjawab pertanyaan pengguna. *RAG* telah terbukti meningkatkan akurasi, keandalan, dan transparasi dari sistem tanya jawab berbasis *LLM* [8] [9].

1.2 Rumusan Masalah Penelitian

Dengan latar belakang yang telah diuraikan, pernyataan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana mengimplementasikan *Retrieval-Augmented Generation (RAG)* ke dalam sistem *chatbot* LLM untuk mengurangi *hallucination*?
2. Bagaimana evaluasi kinerja *chatbot* yang diintegrasikan *Retrieval-Augmented Generation (RAG)* untuk mengurangi *hallucination*?

1.3 Tujuan Penelitian

Berdasarkan pernyataan masalah yang telah disebutkan di atas, adapun tujuan dari penelitian ini adalah:

1. Mengiplementasikan *Retrieval-Augmented Generation (RAG)* ke dalam sistem *chatbot* untuk mengurangi *hallucination*.
2. Mengetahui evaluasi kinerja chatbot yang diintegrasikan *Retrieval-Augmented Generation (RAG)* dalam mengurangi *hallucination*.

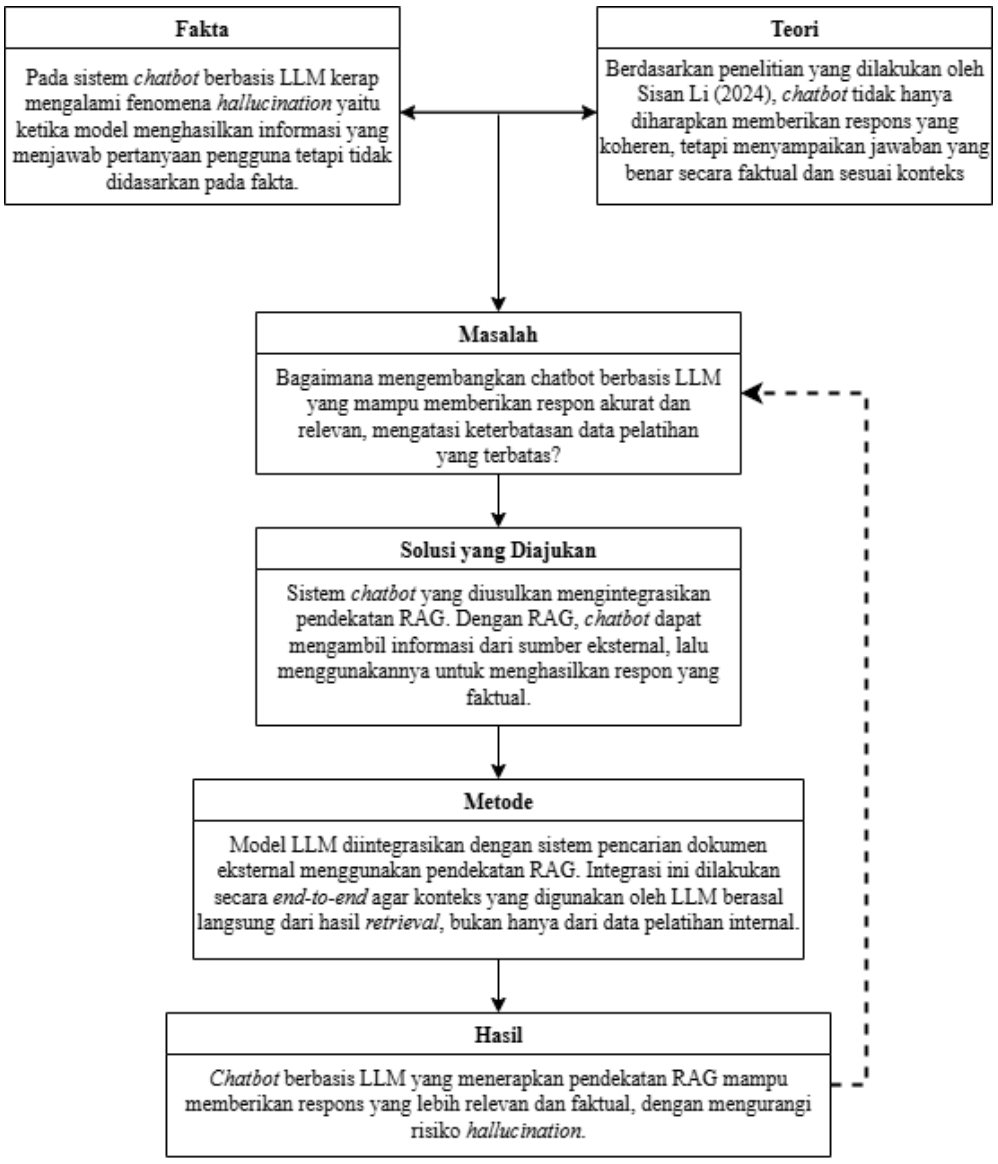
1.4 Batasan Masalah Penelitian

Untuk menjaga fokus penelitian agar tidak terlalu meluas dari pembahasan yang dimaksud, penelitian ini membatasi ruang lingkupnya. Oleh karena itu, batasan masalah dalam penelitian ini meliputi:

1. *Chatbot* yang digunakan mengenai *web programming* menggunakan model *fakhrirudin/llama3-finetuned-full* [10].
2. Dataset untuk korpus RAG dibatasi pada hasil *scraping* dari situs web yang relevan dan dapat diakses secara publik.
3. Dataset untuk *fine-tuning* yang digunakan berjumlah 1000 dataset dari total dataset 59022.
4. Sistem *chatbot* yang dikembangkan dalam penelitian ini hanya difokuskan pada topik seputar pemrograman web.

1.5 Kerangka Pemikiran Penelitian

Kerangka pemikiran merupakan representasi konseptual yang menjelaskan hubungan antar komponen atau variabel utama dalam suatu penelitian. Berikut adalah kerangka pemikiran dari penelitian ini:



Gambar 1. 1 Kerangka Pemikiran

Pada gambar 1.1 menggambarkan alur pemikiran penelitian yang dimulai dari permasalahan umum pada sistem *chatbot* berbasis *Large Language Models* (LLM), yaitu kecenderungan model menghasilkan respons yang tidak didasarkan pada fakta atau informasi yang akurat. Fenomena ini dikenal sebagai *hallucination*, di mana model menjawab pertanyaan pengguna secara meyakinkan, tetapi isi

jawabannya tidak memiliki landasan data yang valid. Hal ini terjadi karena LLM hanya mengandalkan pengetahuan yang diperoleh selama proses pelatihan, tanpa kemampuan untuk mengakses informasi terbaru atau eksternal.

Permasalahan ini diperkuat oleh teori yang dikemukakan dalam penelitian Sisan Li (2024), yang menyatakan bahwa sistem *chatbot* seharusnya tidak hanya memberikan respons yang koheren, tetapi juga harus menyampaikan informasi yang benar secara faktual dan sesuai konteks. Berdasarkan permasalahan tersebut, penelitian ini mengajukan pertanyaan utama, yaitu bagaimana merancang dan mengembangkan sistem *chatbot* berbasis LLM yang mampu memberikan respons yang akurat dan relevan, sekaligus mengatasi keterbatasan data pelatihan yang bersifat statis. Untuk menjawab permasalahan ini, peneliti mengusulkan solusi berupa integrasi pendekatan *Retrieval-Augmented Generation* (RAG). Pendekatan RAG memungkinkan *chatbot* untuk mengambil informasi dari sumber eksternal secara dinamis, kemudian menggunakannya sebagai konteks dalam menghasilkan jawaban.

Metode yang digunakan dalam penelitian ini adalah mengintegrasikan model LLM dengan sistem pencarian dokumen eksternal menggunakan pendekatan RAG secara *end-to-end*. Dengan integrasi ini, konteks yang digunakan oleh model dalam merespons pertanyaan pengguna berasal langsung dari hasil *retrieval*, bukan semata-mata dari parameter model yang dilatih sebelumnya. Pendekatan ini bertujuan untuk memperkaya pengetahuan sistem dan meningkatkan akurasi respons yang dihasilkan.

Dengan diterapkannya pendekatan tersebut, sistem *chatbot* yang dikembangkan diharapkan mampu menghasilkan jawaban yang lebih faktual, relevan, dan kontekstual. Selain itu, penggunaan informasi eksternal juga diharapkan dapat mengurangi risiko *hallucination* serta meningkatkan kualitas interaksi antara sistem dan pengguna.

1.6 Sistematika Penulisan

Berikut adalah sistematika penulisan pada penelitian ini yang disusun secara terstruktur:

BAB I PENDAHULUAN

Pada Bab I yaitu pendahuluan berisi beberapa bahasan seperti latar belakang dari penelitian ini, lalu rumusan masalah pada penelitian, menentukan tujuan serta manfaat apa yang terdapat dalam penelitian, serta membatasi permasalahan pada penelitian ini, tak lupa ada pula kerangka pemikiran, dan sistematikan penulisan yang ditulis secara terstruktur.

BAB II STUDI PUSTAKA

Pada Bab II studi pustaka berisi tentang landasan teori yang mendukung penelitian sehingga menjadi terarah.

BAB III METODOLOGI PENELITIAN

Pada Bab III metodologi penelitian berisikan tentan uraian bagaimana sistem dirancang lalu dibuat dan dimulai dari pemahaman data, pengumpulan data dan proses pengolahan data dari penelitian.

BAB IV HASIL DAN PEMBAHASAN

Pada Bab IV hasil dan pembahasan berisikan tentang hasil dari sistem yang telah dirancang dan dibangun yang nantinya akan dievaluasi.

BAB V PENUTUP

Pada Bab V penutup merupakan tahapan akhir yang berisi tentang kesimpulan dari penelitian ini.

DAFTAR PUSTAKA

Pada daftar pustaka berisi tentang sumber-sumber tertulis yang dipakai dan dijadikan acuan dalam penelitian ini.

LAMPIRAN

Dokumen-dokumen tambahan yang digunakan dalam proses penyusunan dan perancangan penelitian dimuat pada bagian lampiran.