BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Solusi

Penulis menemukan beberapa penelitian terdahulu terkait dengan *chatbot*, dan berbasis LLM + RAG.

2.1.1 Evaluation of a Retrieval-Augmented Generation-Powered Chatbot for Pre-CT Informed Consent: A Prospective Comparative Study

Penelitian yang dilakukan oleh Park et al. ini mengevaluasi keberhasilan implementasi chatbot berbasis LLM + RAG dalam memberikan informasi medis terkait prosedur CT scan kepada pasien. Hasilnya menunjukkan bahwa chatbot mampu menyampaikan informasi yang setara dengan konsultasi langsung dari tenaga medis. Pengukuran dilakukan dengan skala Likert terhadap pemahaman dan kejelasan informasi yang diperoleh pasien. Tidak ditemukan perbedaan signifikan antara chatbot dan dokter ($p > 0,05$), menunjukkan efektivitas chatbot dalam komunikasi berbasis dokumen. Pendekatan RAG memberikan keunggulan karena mampu mengambil data dari dokumen resmi dan menghasilkan jawaban natural melalui LLM.

Dari studi ini, penulis menyimpulkan bahwa LLM + RAG merupakan solusi untuk pengembangan chatbot untuk para petani, karena mampu memberikan informasi yang tepat kepada para petani. [8]

2.1.2 Increasing customer service efficiency through artificial intelligence chatbot

Penelitian dengan judul "Increasing customer service efficiency through artificial intelligence chatbot" yang dilakukan oleh Ivan Martins De Andrade dan Cleonir Tumelero [9]. Penelitian ini membahas penerapan *chatbot* berbasis AI dalam meningkatkan efisiensi layanan pelanggan. *Chatbot* dapat mempercepat layanan, meningkatkan aksesibiltias, dan mengurangi beban kerja agen manusia dengan menangani pertanyaan

berulang. Selain itu, *chatbot* membantu mengurangi antrian, memberikan interaksi yang lebih responsif, dan menyediakan layanan 24 jam tanpa ketergantungan pada agen manusia.

Dari penelitian ini, penulis akan menggunakan *chatbot* dikarenakan dapat meningkatkan efisiensi komunikasi antara pengguna dan sistem. *Chatbot* dapat membantu dan memberikan informasi yang cepat dan akurat. *Chatbot* dapat memastikan pengalaman pengguna yang lebih baik.

2.1.3 Retrieval-Augmented Generation: A Survey

Dalam "Retrieval-Augmented Generation: A Survey" membahas secara mendalam bagaimana pendekatan RAG (Retrieval-Augmented Generation) meningkatkan kinerja model bahasa besar (LLM). Dalam studi tersebut, RAG terbukti dapat meningkatkan kualitas jawaban dengan menggabungkan kekuatan pemahaman bahasa alami dari LLM dengan kemampuan pencarian informasi eksternal secara real-time. Dalam penerapan RAG untuk chatbot MySalak, evaluasi performa sistem juga harus dilakukan secara objektif.

Salah satu pendekatan yang dapat digunakan adalah framework RAGAS (Retrieval-Augmented Generation Assessment Suite), yang menyediakan metrik evaluasi seperti faithfulness dan context precision. Metrik faithfulness mengukur sejauh mana jawaban sesuai dengan konteks yang tersedia, sedangkan context precision menilai apakah konteks yang digunakan benar-benar relevan terhadap pertanyaan. Penggunaan RAGAS akan memastikan bahwa jawaban yang diberikan oleh chatbot tidak hanya akurat, tetapi juga berdasarkan pada dokumen yang tepat dan valid. [10]

2.2 Tinjauan Teori

2.2.1 Chatbot

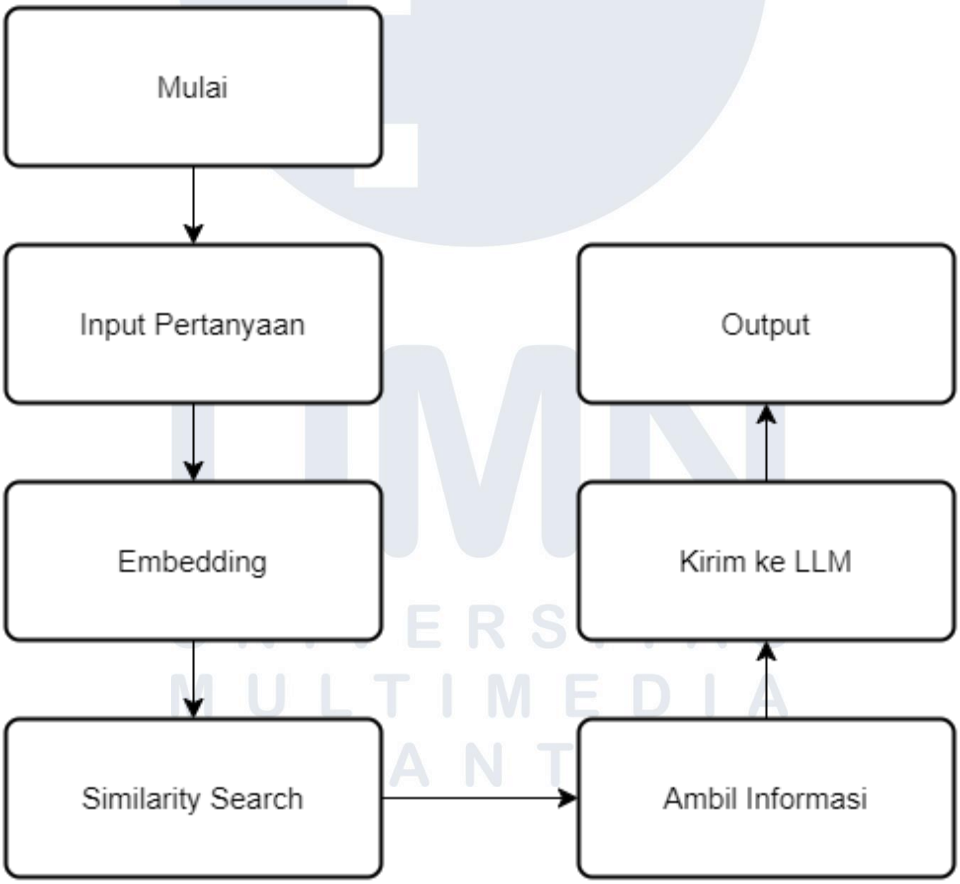
Chatbot adalah program AI yang dirancang untuk simulasikan percakapan dengan pengguna manusia melalui teks ataupun suara. *Chatbot* menggunakan

Natural Language Processing (NLP) dan analisis sentimen untuk memahami dan merespons bahasa manusia. Mereka juga dikenal sebagai *aritifical conversation entities, digital assitants,* atau *interactive agents* [11].

2.2.2 Artificial Intelligence

Artificial Intelligence atau yang biasa disebut dengan AI adalah studi tentan bagaimana membuat sebuah komputer dapat melakukan tugas-tugas yang biasanya dilakukan oleh kecerdasan manusia. AI dirancang untuk meniru beberapa fungsi otak manusia. Seperti pemahaman bahasa, pengetahuan, pemecahan masalah, dan lain-lain. AI memungkinkan komputer untuk menerima pengetahuan melalui manusia dan menggunakannya unutk simulasikan proses berpikir manusia dalam memecahkan masalah. [12]

2.2.3 Retrieval Augmented Generation (RAG)



Gambar 2.1 Flowchart RAG

Retrieval-Augmented Generation (RAG) merupakan pendekatan terkini dalam pengembangan chatbot berbasis Artifical Intelligence, yang menggabungkan Large Language Model dengan sistem retriever. Arsitektur RAG bertujuan untuk mengatasi keterbatasan LLM dalam memberikan jawaban akurat atas pertanyaan yang bersifat domain-spesifik, di mana model sering mengalami fenomena hallucination atau menghasilkan informasi yang tidak tepat.

RAG bekerja dengan cara mengambil potongan informasi (chunks) dari knowledge base eksternal yang relevan terhadap pertanyaan pengguna, kemudian menyisipkan informasi tersebut ke dalam prompt sebelum diproses oleh LLM untuk menghasilkan jawaban yang lebih baik.

Keunggulan utama dari pendekatan ini adalah kemampuannya dalam meningkatkan keakuratan jawaban, mempercepat proses pencarian informasi, dan mengurangi ketergantungan pada retraining model ketika ada pembaruan data. Sistem chatbot yang dibangun dengan pendekatan RAG sangat cocok untuk diterapkan dalam konteks menjadi customer service untuk para petani [13]

2.2.4 MySalak

MySalak adalah aplikasi yang bertujuan untuk membantu petani salak dalam pengendalian hama dengan memanfaatkan AI dan IoT. Aplikasi ini menawarkan fitur-fitur seperti perhitungan otomatis jumlah lalat, peta distribusi hama, prediksi cuaca, serta artikel informatif. Selain itu, MySalak terintegrasi dengan perangkat keras *MySalak Node* yang berfungsi sebagai alat pemantauan lingkungan berbasis data, mendukung petani dalam mengelola tanaman salak secara efektif [1].

2.2.5 Large Language Model (LLM)

Large Language Model (LLM) adalah jenis kecerdasan buatan yang dibuat untuk bisa memahami dan menghasilkan teks seperti manusia. Model ini dilatih dengan membaca banyak sekali teks dari internet, buku, artikel, dan sumber lainnya. Karena dilatih dengan data yang sangat besar, LLM bisa

menjawab pertanyaan, membuat rangkuman, menerjemahkan bahasa, atau membantu menulis teks secara otomatis. LLM menggunakan teknologi bernama transformer yang membuatnya mampu mengerti konteks dan pola dalam kalimat. Contoh LLM yang terkenal adalah GPT-3, GPT-4, dan lainnya. Keunggulan utama LLM adalah bisa digunakan untuk banyak tugas tanpa harus dilatih ulang dari awal, sehingga sangat fleksibel dan praktis untuk berbagai kebutuhan. [14]