PROPOSAL TUGAS AKHIR

IMPLEMENTASI GRAPH-BASED RETRIVAL AUGMENTED GENERATION (GraphRAG) PADA LARGE LANGUAGE MODEL (LLM) UNTUK PENGEMBANGAN CHATBOT VIRTUAL ASSISTANT UNIVERSITAS SUMATERA UTARA

DHANI DWI SEPTIAN BANGUN

211402120

PROGRAM STUDI S1 TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS SUMATERA UTARA

MEDAN

2025

BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi informasi berkembang pesat dan telah membawa dampak yang signifikan di berbagai bidang, termasuk bidang pendidikan. Teknologi membantu pekerjaan manusia dalam berbagai bidang menjadi lebih cepat terselesaikan. Di era digital ini, universitas dituntut untuk menyediakan layanan informasi yang cepat, akurat, dan efisien guna memenuhi kebutuhan mahasiswa dan masyarakat luas (Khan, 2018). Universitas Sumatera Utara (USU) memiliki berbagai layanan berbasis website seperti portal akademik, situs resmi universitas, situs fakultas, dan layanan lainnya. Namun, banyaknya website yang tersedia sering kali membuat mahasiswa, dosen, maupun staf kesulitan mengakses informasi yang mereka butuhkan secara cepat dan efisien. Pengguna harus mengunjungi beberapa situs atau aplikasi berbeda untuk menemukan informasi spesifik, yang dapat mengakibatkan kebingungan dan pemborosan waktu. Untuk mengatasi kendala ini, Universitas Sumatera Utara perlu mengembangkan solusi yang lebih terintegrasi dan ramah pengguna.

Salah satu alternatif yang dapat diterapkan adalah pengembangan virtual assistant berbasis kecerdasan buatan menggunakan chatbot. Chatbot adalah program komputer berbasis kecerdasan buatan yang dapat melakukan percakapan melalui audio atau teks (Haristiani, 2019). Chatbot memiliki kemampuan untuk memahami permintaan pengguna dan memberikan respons yang relevan dalam waktu singkat, sehingga menghemat waktu dibandingkan pencarian manual melalui situs web atau mesin pencari seperti Google. Dengan kemampuan chatbot untuk merespons pertanyaan dengan cepat dan tepat, diharapkan dapat mengurangibeban kerja staf administrasi serta meningkatkan pengalaman pengguna dalam mengakses informasi akademik (Hussain et al., 2021). Salah satu penerapan Chatbot yaitu menggunakan Large Language Models (LLM). LLM adalah salah

satu jenis model AI yang dapat memproses dan menghasilkan teks bahasa. Model pada LLM umumnya dilatih dengan menggunakan sejumlah besar data teks dan menggunakan teknik pembelajaran mendalam atau deep learning untuk mempelajari pola dan struktur bahasa (M.U. Hadi et al, 2021). Kehadiran LLM sendiri menjadi sebuah fenomena luar biasa pada beberapa tahun terakhir ini, tepatnya setelah ke hadiran ChatGPT yang dirilis oleh perusahaan OpenAI pada bulan November 2022 yang lalu. Hanya dalam kurun waktu dua bulan setelah perilisannya, pada Januari 2023 jumlah pengguna ChatGPT telah mencapai 100 juta pengguna di seluruh dunia (Hu & Krystal, 2023).

Hal ini mengindikasikan bahwa chatbot dapat menjadi solusi efektif untuk meningkatkan aksesibilitas informasi, khususnya di lingkungan universitas. Chatbot semakin popular digunakan dalam bidang Pendidikan karena dapat meningkatkan interaksi antara institusi dan pengguna. Dalam studi yang dilakukan oleh Fatima et al. (2020), ditemukan bahwa chatbot dapat membantu mengurangi waktu tunggu untuk mendapatkan informasi dan menyediakan dukungan 24/7 kepada pengguna. Hal inisangat penting dalam konteks akademik, di mana informasi sering kali dibutuhkan di luar jam kerja.

Dalam penelitian ini, metode Graph-Based Retrieval Augmented Generation (GraphRAG) diimplementasikan untuk mendukung performa chatbot berbasis LLM. Selain itu, pendekatan GraphRAG tidak hanya meningkatkan akurasi jawaban, tetapi juga memastikan bahwa informasi yang disampaikan selalu relevan dan terkini. Hal ini menjadi penting dalam lingkungan universitas, di mana kebutuhan informasi dapat sangat dinamis, mencakup jadwal akademik, pengumuman, prosedur administratif, hingga data penelitian. Dengan menerapkan GraphRAG, chatbot mampu mengakses dan menghubungkan berbagai sumber data internal Universitas Sumatera Utara secara efektif, sehingga memberikan pengalaman yang lebih personal dan interaktif bagi penggunanya.

Berdasarkan permasalahan yang sudah diuraikan sebelumnya, maka sangat diperlukan solusi untuk mengatasinya. Salah satu cara untuk meminimalisir waktu dalam penyampaian informasi akademik di Universitas Sumatera Utara dengan

penerapan chatbot, diharapkan dapat mengatasi masalah keterlambatan dalam mengakses informasi akademik.

Terdapat penelitian terkait yang dilakukan oleh (Galstyan & Martirosyan, 2024) yang berjudul "SmartAdvisor University Chatbot Spring 2024". Penelitian ini memanfaatkan kemampuan model pembuatan teks untuk menghadirkan chatbot penasihat akademik yang inovatif dan personal. Tujuan utama penelitian ini untuk mengembangkan sumber daya menyeluruh yang menggabungkan pengetahuan tentang universitas program studi untuk mahasiswa. Penelitian ini memanfaatkan basis data vektor bersama dengan kerangka kerja Retrival-Augmented Generation (RAG).

Penelitian lainnya dilakukan oleh (Rachmat & Kesuma, 2024) yang berjudul Implementasi Large Language Models Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android. Penelitian ini dilakukan untuk memahami bagaimana implementasi API key Gemini. API key Gemini dapat dilakukan pada aplikasi AI, khususnya pada aplikasi chatbot, dan apakah aplikasi chatbot yang dibangun menggunakan Gemini dapat berfungsi sesuai dengan fungsionalitas Gemini LLM itu sendiri. Setelah aplikasi chatbot yang dimaksud selesai dibuat dan diuji coba, didapatkan bahwa aplikasi chatbot yang dihasilkan aplikasi chatbot yang dihasilkan berfungsi sesuai dengan yang diharapkan dan mampu memanfaatkan fungsionalitas LLM Gemini.

Penelitian lainnya dilakukan oleh (Odede Ingo Frommholz JAOdede, 2024) yang berjudul JayBot – Aiding University Students and Admission with an LLM-based Chatbot. Penelitian ini menyajikan JayBot yaitu sistem chatbot berbasis LLM yang bertujuan untuk meningkatkan pengalaman pengguna dosen, staf, calon mahasiswa dan mahasiswa di sebuah universitas di inggris.

Penelitian lainnya dilakukan oleh (Radhakrishnan & Dias, 2023) yang berjudul Use of a ChatBot-Based Advising System for the Higher-Education System. Penelitian ini memperkenalkan mekanisme yang layak untuk sistem pemberian saran berbasis chatbot. penelitian ini menggunakan Large Language

Model (LLM) yang bersifat opensource yang dikombinasikan dengan basis pengetahuan khusus untuk mengatasi aspek-aspek penting yang diperlukan untuk sistem pemberian saran berbasis chatbot, seperti personalisasi, memori percakapan, dan kemudahan pemeliharaan. Sistem yang diusulkan menunjukkan tingkat akurasi respons sebesar 89% yang membuktikan bahwa pendekatan baru dari arsitektur berbasis komponen ini unggul dalam hal kinerja jika dibandingkan dengan pendekatan serupa.

Berdasarkan dari latar belakang dan beberapa penelitian terdahulu yang telah diuraikan di atas, dapat ditarik kesimpulan bahwa sistem Natural Language Processing merupakan suatu solusi yang tepat. Maka penulis mengajukan untuk melakukan penelitian mengenai pembuatan system **"Implementasi Graph-Based Retrieval Augmented Generation (GraphRAG) pada Large Language Model (LLM) untuk Pengembangan Chatbot Virtual Assistant Universitas Sumatera Utara".**

1.2 Rumusan Masalah

Universitas Sumatera Utara (USU) menyediakan berbagai layanan berbasis website seperti portal akademik, situs resmi universitas, dan situs fakultas. Namun, dengan banyaknya situs yang tersedia, mahasiswa mengalami kesulitan dalam mengakses informasi dengan cepat dan efisien. Proses pencarian informasi yang tersebar di berbagai platform dapat membingungkan pengguna dan membuang waktu yang seharusnya bisa digunakan untuk tujuan lainnya. Oleh karena itu dibutuhkan sebuah sistem chatbot berbasis kecerdasan buatan (AI) yang dapat menyederhanakan proses pencarian dengan memberikan jawaban yang cepat, relevan, dan mudah diakses oleh pengguna. Penerapan teknologi ini dalam chatbot di lingkungan universitas dapat meningkatkan pengalaman pengguna dalam mengakses informasi akademik dan administratif secara lebih cepat dan akurat.

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan chatbot virtual assistant Universitas Sumatera Utara (USU) dengan mengimplementasikan Graph-Based Retrival Augmented Generation (GraphRAG) pada Large Language Model (LLM).

1.4 Batasan Masalah

Agar penelitian ini fokus yang jelas dan tidak menyimpang dari tujuan serta mecegah terlalu luasnya ruang lingkup pembahasan dalam penelitian ini, oleh karena itu diperlukan beberapa batasan masalah, yaitu:

1. Seluruh informasi yang berada pada website usu.ac.id dan sub domain usu.ac.id
2. Data yang digunakan untuk retrieval dalam knowledge based akan dikonversi ke dalam format file berekstensi .pdf.
3. Penelitian ini berfokus pada penerapan teknologi Graph Based Retrieval-Augmented Generation (GraphRAG) dengan dukungan LangChain.
4. Data yang mengandung gambar dan link URL dihapus selama proses preprocessing.
5. User dari chatbot ini berfokus pada mahasiswa Universitas Sumatera Utara.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini, yaitu:

1. Memudahkan pencarian informasi mengenai Universitas Sumatera Utara
2. Mendukung transformasi digital Universitas Sumatera Utara
3. Menjadi refrensi untuk penelitian selanjutnya
4. Menguji performa GraphRAG daam LLM dalam pengembangan chatbot Universitas Sumatera

1.6 Metodologi Penelitian

Adapun tahapan-tahapan penelitian yang akan dilakukan, yaitu:

1.6.1 Studi Literatur

Pada tahap awal ini, penulis melakukan pengumpulan setiap informasi yang dibutuhkan dari sumber-sumber terpercaya seperti jurnal, artikel, buku, skripsi, dan informasi valid lainnya terkait chatbot, Large Language Model (LLM), dan Graph-Based Retrieval Augmented Generation (GraphRAG).

1.6.2 Analisis Permasalahan

Setelah memperoleh informasi yang telah dikumpulkan, penulis melakukan analisis sesuai dengan yang dibutuhkan dalam melakukan pengembangan chatbot virtual assistant Universitas Sumatera Utara (USU) dengan mengimplementasikan Graph-Based Retrieval Augmented Generation (GraphRAG) dan Large Language Model (LLM).

1.6.3 Perancangan Sistem

Melakukan perancangan sistem, mulai dari arsitektur umum dan mengumpulkan data untuk mengembangkan chatbot virtual assistant Universitas Sumatera Utara menggunakan Graph-based Retrieval Augmented Generation (GraphRAG) yang diintegrasikan dengan Large Language Models (LLM).

1.6.4 Implementasi Sistem

Pada tahap ini dilakukan penerapan terhadap perancangan sistem yang telah dibuat sebelumnya yang bertujuan untuk mencapai tujuan sesuai dengan konteks dan ruang lingkup penelitian.

1.6.5 Pengujian Sistem

Pada tahap ini sistem yang telah dibangun akan melakukan tahap pengujian untuk melihat kinerja sistem yang telah dirancang dan memastikan bahwa sistem dapat berfungsi dengan baik.

1.6.6 Penyusunan Laporan

Tahap ini merupakan tahap akhir dari seluruh rangkaian proses penelitian dimana seluruh proses penelitian disusun menjadi sebuah laporan yang disertai dengan dokumentasi yang menunjukkan hasil akhir dari penelitian.

1.7 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini akan dibagi menjadi lima bagian, yaitu:

BAB I PENDAHULUAN

Bab ini mencakup pendahuluan penelitian yang meliputi latar belakang dilakukannya penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bagian ini memuat teori-teori yang relevan sebagai dasar dan pendukung penelitian yang dilakukan. Teori-teori ini terkait dengan pengembangan chatbot Universitas Sumatera Utara berbasis Large Language Model (LLM) yang memanfaatkan Graph-Based Retrieval Augmented Generation (GraphRAG).

BAB III ANALISIS DAN PERANACANGAN SISTEM

Bab tiga menjelaskan analisis sistem yang mencakup arsitektur umum, tahapan pengembangan, dan perancangan chatbot. Pembahasannya meliputi chatbot berbasis Large Language Model (LLM) yang memanfaatkan GraphRAG.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Tahap-tahap implementasi sistem chatbot Universitas Sumatera Utara (USU), mulai dari penerapan LLM dan integrasi dengan RAG hingga pengujian sistem. Evaluasi dilakukan untuk mengukur akurasi respons chatbot, kepuasan pengguna, serta kemampuan sistem dalam menangani keluhan pelanggan.

BAB V KESIMPULAN DAN SARAN

Kesimpulan dari hasil penelitian yang telah dilakukan, mencakup efektivitas penerapan LLM dan RAG dalam chatbot Universitas Sumatera Utara (USU). Selain itu, diberikan saran untuk pengembangan lebih lanjut agar chatbot dapat lebih optimal di masa mendatang.

Sistematika ini dirancang untuk memberikan struktur yang jelas dan sistematis dalam memaparkan penelitian mengenai chatbot berbasis LLM dan RAG pada chatbot Universitas Sumatera Utara (USU).

BAB II

LANDASAN TEORI

2.1 Chatbot

Chatbot adalah program komputer yang dirancang untuk berinteraksi dengan manusia menggunakan bahasa alami. Teknologi ini banyak dimanfaatkan untuk berbagai keperluan, seperti layanan bantuan daring, layanan personal, dan penyebaran informasi. Chatbot merupakan salah satu aplikasi dari Natural Language Processing (NLP), yang merupakan cabang dari kecerdasan buatan (Artificial Intelligence). Dengan memanfaatkan Artificial Intelligence dan NLP, chatbot dapat memahami dan menganalisis teks dari pengguna, lalu memberikan respons otomatis sesuai dengan pertanyaan atau perintah yang diberikan melalui pesan (’Wicaksana et al., 2024).

Chatbot dapat berfungsi sebagai agen percakapan yang mampu membantu atau menggantikan peran seorang konsultan. Chatbot dilengkapi dengan basis pengetahuan yang memungkinkan untuk berinteraksi dan berdialog dengan manusia. Dari segi fungsionalitas, chatbot dapat beroperasi layaknya customer service dalam bentuk sistem aplikasi.

2.2 Virtual Assistant

Virtual assistant adalah sebuah sistem yang bertindak seperti asisten tapi segala kinerjanya dilakukan secara otomatis sesuai program, salah satu bentuk virtual assistant adalah chatbot (Bariyah & Imania, 2022). Chatbot asisten virtual adalah program komputer yang dirancang untuk mensimulasikan percakapan manusia guna memberikan layanan atau informasi secara otomatis. Teknologi ini memanfaatkan pemrosesan bahasa alami dan pembelajaran mesin untuk memahami serta merespons pertanyaan pengguna secara efektif. Dalam dunia pendidikan, chatbot digunakan sebagai asisten virtual untuk menjawab pertanyaan mahasiswa, memberikan informasi akademik, serta membantu proses pembelajaran, sebagaimana diteliti oleh Damayanti dan Nuzuli (2023) yang menunjukkan implementasi chatbot sebagai asisten virtual di lingkungan kampus.

2.3 Natural Language Processing (NLP)

Bahasa merupakan sistem aturan atau koleksi simbol yang dikompilasi dan digunakan untuk menyampaikan informasi. Tetapi tidak semua orang mempunyai waktu dan kemampuan dalam mempelajari bahasa pemrograman. Oleh karena itu, Pemrosesan Bahasa Alami (Natural Language Processing) sering menjadi solusi untuk mempelajari bahasa pemrograman secara mendalam. Pemrosesan Bahasa Alami adalah subbidang ilmu komputer dan kecerdasan buatan (artificial intelligence) yang berfokus pada kemampuan mesin untuk memahami dan memproses bahasa manusia dengan cara yang efektif dan akurat. Teknik NLP (Natural Language Processing) digunakan untuk menyelesaikan berbagai tugas yang berhubungan dengan bahasa, seperti pengenalan suara, penerjemahan bahasa, analisis sentimen, chatbot, dan banyak lagi (Rantung, 2023) .

NLP (Natural Language Processing) bergantung pada kombinasi linguistik, statistik, dan pembelajaran mesin untuk memungkinkan komputer memahami bahasa manusia dan menghasilkan respons yang sesuai dan relevan dengan input yang mereka terima (Rantung, 2023). Secara umum, NLP (Natural Language Processing) dibagi menjadi dua aspek utama, yaitu Natural Language Understanding, tentang bahasa yang memiliki fokus pada pemahaman teks, dan Natural Language Generation, yang berguna untuk menghasilkan teks berdasarkan kebutuhan pengguna. Kedua aspek terus dikembangkan untuk mendukung komunikasi yang lebih efisien antara manusia dan mesin (Khurana et al., 2023).

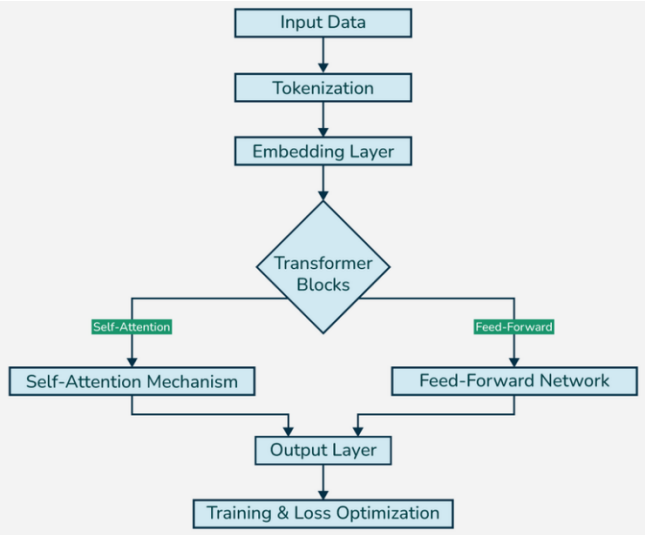
2.4 Large Language Model (LLM)

LLM telah menjadi inovasi besar yang mengubah cara memproses, menghasilkan, dan memahami bahasa manusia. Model ini dilatih menggunakan kumpulan data teks yang sangat luas, mencakup volume data bahasa yang beragam, sehingga mampu menyelesaikan tugas-tugas yang membutuhkan pemahaman konteks dan menghasilkan respons yang relevan serta bermakna (Shafee et al., 2024). Generasi terbaru dari model bahasa telah diterapkan secara luas dalam berbagai bidang.

Sebagai contoh, LLM telah dimanfaatkan untuk mendukung pemecahan masalah di bidang matematika, fisika, dan kimia (Arora et al., 2023).

Penelitian oleh Arora dan Singh (2023), menunjukkan bahwa LLM memiliki kemampuan yang kuat dalam penalaran analogis, tetapi kesulitan dalam tes penalaran spasial. Tidak hanya itu, berdasarkan penelitian oleh (Chang et al., 2024), Large Language Models (LLM) memiliki kemampuan untuk menghasilkan teks yang tampak koheren dan faktual. Namun, hasil yang dihasilkan terkadang mengandung ketidakakuratan atau pernyataan yang tidak berbasis pada fakta, sebuah fenomena yang dikenal sebagai halusinasi. Selain itu, LLM juga dapat mencerminkan bias sosial dan menghasilkan toksisitas selama proses pembuatan teks, yang dapat menyebabkan output yang bias.

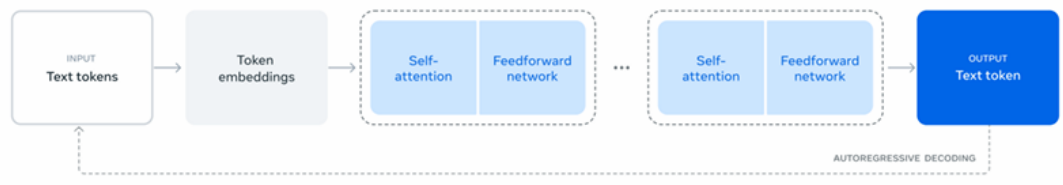
Berdasarkan penelitian sebelumnya, meskipun LLM menunjukkan potensi besar dalam berbagai bidang, model ini masih menghadapi sejumlah tantangan. Selain kemampuan luar biasa dalam penalaran analogis, terdapat kelemahan yang perlu diperhatikan, seperti kesulitan dalam memahami aspek-aspek tertentu, termasuk penalaran spasial dan ketergantungan pada data pelatihan. Adapun arsitektur LLM dapat dilihat pada gambar 2.1 berikut ini.



Gambar 2.1 Arsitektur LLM
(Sumber : LLM Architecture: Exploring the Technical Architecture Behind Large Language Models - GeeksforGeeks)

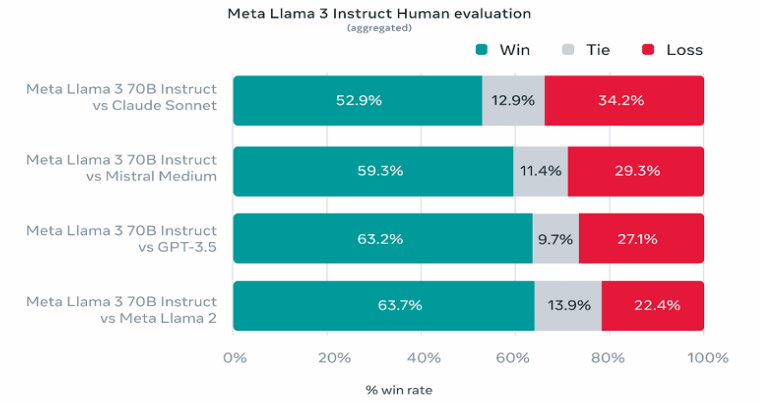
2.5 Llama-3

Llama-3 adalah model bahasa yang mendukung kemampuan multibahasa, pemrograman, penalaran, dan penggunaan alat. Melalui evaluasi empiris yang ekstensif, ditemukan bahwa Llama-3 mampu memberikan kualitas yang sebanding dengan model bahasa terkemuka seperti GPT-4 dalam berbagai tugas (Grattafiori et al., 2024). Arsitektur Llama-3 dapat dilihat pada gambar 2.2 berikut ini.



Gambar 2.2 Ilustrasi keseluruhan arsitektur dan pelatihan Llama 3

Model Llama-3 dengan parameter 8 miliar (8B) dan 70 miliar (70B) merupakan peningkatan besar dibandingkan Llama-2 . Dampak perbaikan dalam proses pra-pelatihan dan pasca-pelatihan, model ini yang telah dilatih dan disesuaikan dengan instruksi menjadi salah satu yang terbaik pada skala parameter 8B dan 70B. Peningkatan dalam prosedur pasca-pelatihan secara signifikan mengurangi tingkat penolakan palsu, meningkatkan penyelarasan, serta memperluas keragaman respons (Meta AI, 2024). Perbandingan LLama-3 dengan model LLM lainnya dapat dilihat pada gambar 2.3 berikut ini.



Gambar 2.3 Hasil evaluasi manusia dibandingkan dengan model lain seperti Claude Sonnet, Mistral Medium, dan GPT-3.5
(Sumber: Meta AI, 2024)

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

Gambar 2.4 Perbandingan Model Large Language Model
(Sumber: https://arxiv.org/pdf/2407.21783)

Selain itu, berdasarkan penelitian Budey et al., (2024), Llama 3 memiliki beberapa keunggulan seperti performa yang kompetitif dalam berbagai. Pada benchmark MMLU, Llama 3 70B mencapai skor 83.6, mengungguli model seperti Mistral 7B yang hanya memperoleh 61.1 dan Gemma 2 9B dengan 72.3. Selain itu, pada evaluasi HumanEval untuk pengkodean, Llama 3 70B mencatatkan skor 80.5, yang mendekati performa model closed-source GPT-3.5 Turbo (68.0) dan GPT-4 (86.0).

Dalam kategori matematika, Llama 3 70B juga menunjukkan keunggulan signifikan, terutama pada benchmark GSM8K, dengan skor 95.1, melampaui performa GPT-3.5 Turbo (88.2) dan Nemotron 43B (92.3). Keunggulan ini semakin terlihat pada kategori ARC Challenge, di mana Llama 3 70B berhasil mencapai skor 94.3, lebih tinggi dibandingkan GPT-3.5 Turbo (88.7) dan model open-source lainnya.

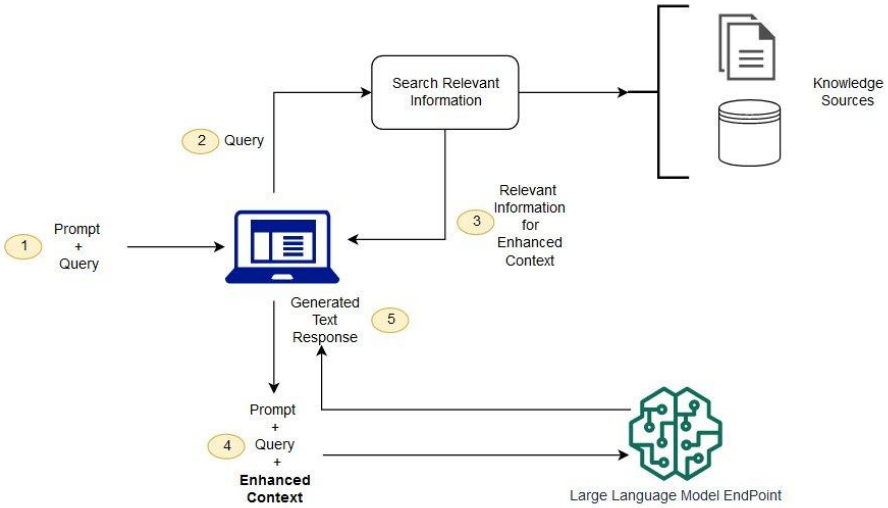
Selain itu, Llama 3 70B juga menunjukkan performa yang baik dalam Multilingual General Language Understanding (MGSM) dengan skor 86.9, menandakan kemampuan yang kuat dalam pemrosesan bahasa multibahasa. Keunggulan ini menunjukkan bahwa Llama 3 dapat menjadi alternatif yang kompetitif bagi model closed-source seperti GPT-4 dan Claude 3.5 Sonnet dalam

berbagai tugas NLP. Informasi detail perbandingan Llama-3 dapat dilihat pada gambar 2.4 diatas.

2.6 Retrival-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) telah menjadi pendekatan inovatif dalam pemrosesan bahasa alami, yang mengintegrasikan keunggulan Large Language Models (LLM) dengan kemampuan untuk mengambil dan menggabungkan informasi relevan dari sumber pengetahuan eksternal. Penelitian telah mengeksplorasi penerapan RAG di berbagai bidang, menunjukkan potensinya dalam meningkatkan performa sistem tanya-jawab dan dialog berbasis AI (Quidwai & Lagana, 2024).

Penelitian oleh Lewis et al., (2020) menjelaskan tentang arsitektur RAG yang yang mengintegrasikan model bahasa pra-terlatih dengan mekanisme pengambilan informasi untuk menghasilkan respons berbasis dokumen yang relevan. Mereka menunjukkan efektivitas RAG dalam berbagai tugas yang membutuhkan pengetahuan mendalam, seperti menjawab pertanyaan domain terbuka dan melakukan verifikasi fakta. Arsitektur umum RAG dapat dilihat pada gambar 2.5 berikut ini.



Gambar 2.5 Arsitektur Retrieval-Augmented Generation (RAG)

(Sumber: https://aws.amazon.com/id/what-is/retrieval-augmented-generation/)

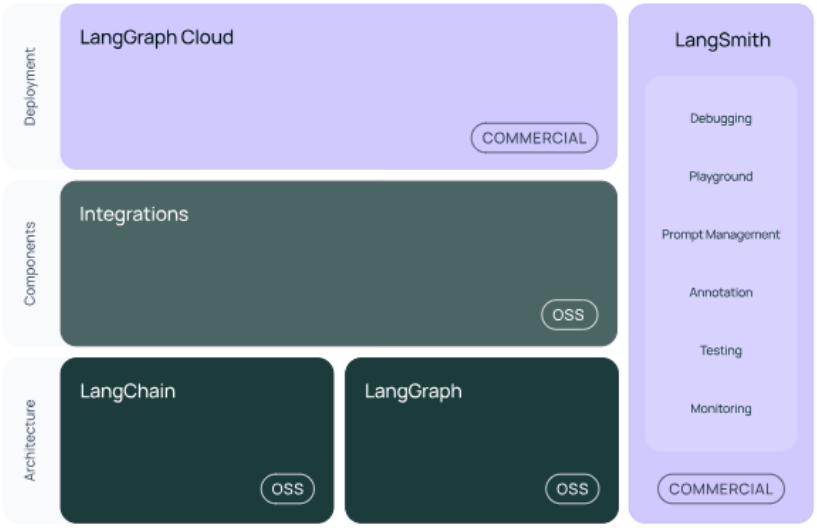
2.7 Few-shot Prompting

Few-shot (FS) Prompting adalah salah satu teknik dari prompt engineering. Metode ini memberikan beberapa contoh tugas selama proses inferensi sebagai referensi, tanpa diperbolehkan melakukan pembaruan bobot. Dalam pendekatan few-shot, model diberikan K contoh yang mencakup konteks beserta hasilnya, diikuti oleh satu contoh terakhir yang hanya berisi konteks, dengan harapan model dapat memprediksi hasil yang sesuai. Biasanya, nilai K berkisar antara 10 hingga 100, sesuai dengan jumlah contoh yang dapat dimasukkan dalam jendela konteks model. Keunggulan utama few-shot learning adalah mengurangi kebutuhan akan data khusus untuk setiap tugas dan menghindari pembelajaran dari distribusi data yang terlalu sempit akibat fine-tuning pada dataset yang sangat spesifik. Namun, kelemahan dari metode ini adalah performanya masih di bawah model fine-tuned yang telah mencapai tingkat keunggulan (state-of-the-art). Selain itu, meskipun kebutuhan akan data tugas khusus berkurang, tetap diperlukan sejumlah kecil data untuk mendapatkan hasil yang optimal (Brown et al., 2020).

2.8 Langchain

LangChain adalah sebuah kerangka kerja yang dirancang untuk mengembangkan aplikasi oleh model bahasa besar (LLM). Ini bertindak sebagai antarmuka umum, membantu dengan integrasi hampir semua LLM ke dalam berbagai aplikasi (Колодяжна, 2024). LangChain memfasilitasi pengembang dalam membuat aplikasi berbasis LLM, dengan fokus pada peningkatan kustomisasi, akurasi, dan relevansi model.

Kemudian, Penelitian yang dilakukan oleh Oughuzan Topsakal dan T. Cetin Akinci (2023) tentang pengembangan aplikasi berbasis Large Language Model (LLM) dengan memanfaatkan LangChain menunjukkan bahwa LangChain berperan sebagai kerangka kerja yang mendukung pengembang dalam mengoptimalkan penggunaan LLM secara efektif, efisien, dan terstruktur. Langchain introduction dapat dilihat pada gambar 2.6 berikut ini.



Gambar 2.6 Langchain Introduction

(Sumber: https://js.langchain.com/docs/introduction)

Tabel 2.1 Penelitian Terdahulu

No.	Penulis	Judul	Tahun	Keterangan
1.	Lilit Galstyan dan Hovhannes Martirosyan	SmartAdvisor University Chatbot Spring 2024	2024	Penelitian ini memanfaatkan kemampuan model pembuatan teks untuk menghadirkan chatbot penasihat akademik yang bertujuan untuk mengembangkan sumber daya menyeluruh yang menggabungkan pengetahuan tentang universitas program studi untuk mahasiswa. Penelitian ini menggunakan basis data vektor bersama dengan kerangka kerja Retrieval-Augmented Generation (RAG), serta mengintegrasikan UI yang ramah pengguna yang membuat interaksi dan navigasi yang sederhana, chatbot dapat menjawab pertanyaan tentang mata kuliah individu dan memberikan informasi berdasarkan semester, mata kuliah semester, deskripsi mata kuliah, dan prasyarat.
2.	Nur Rachmat dan Dorie P. Kesuma	Implementasi *Large Language Models Gemini* Pada Pengembangan Aplikasi *Chatbot* Berbasis Android	2024	Penelitian ini dilakukan untuk memahami bagaimana implementasi API key Gemini. API key Gemini dapat dilakukan pada aplikasi AI, khususnya pada aplikasi chatbot, dan apakah aplikasi chatbot yang dibangun menggunakan Gemini dapat berfungsi sesuai dengan fungsionalitas Gemini LLM itu sendiri. Dalam mengembangkan aplikasi chatbot menggunakan LLM Gemini, penulis menggunakan metodologi pengembangan prototyping, dan aplikasi chatbot yang dikembangkan berjalan pada

				platform Android. Setelah aplikasi chatbot yang dimaksud selesai dibuat dan diuji coba, didapatkan bahwa aplikasi chatbot yang dihasilkan aplikasi chatbot yang dihasilkan berfungsi sesuai dengan yang diharapkan dan mampu memanfaatkan fungsionalitas LLM Gemini.
3.	Leonardo pasquarelli, Charles Koutcheme, Arto Hellas	Comparing the Utility, Preference, and Performance of Course Material Search Functionality and Retrieval-Augmented Generation Large Language Model (RAG-LLM) AI Chatbots in Information-Seeking Tasks	2024	Penelitian ini mengembangkan chatbot AI bertenaga LLM yang menambah jawaban yang dihasilkan dengan informasi dari materi kursus. Penelitian ini menyediakan dukungan yang memadai bagi mahasiswa membutuhkan sumber daya yang besar, terutama mengingat jumlah mahasiswa yang terus bertambah.
4.	Julius Odede, Ingo Frommholz	JayBot – Aiding University Students and Admission with an LLM-based Chatbot	2024	Jaybot merupakan sistem chatbot berbasis LLM yang bertujuan untuk meningkatkan pengalaman pengguna calon mahasiswa dan mahasiswa, fakultas, dan staf di sebuah universitas di Inggris. Tujuannya dari JayBot adalah untuk memberikan informasi kepada pengguna tentang pertanyaan umum mengenai modul kursus, durasi, biaya, persyaratan masuk, kuliah turers, magang, jalur karier, kelayakan kerja kursus dan lainnya aspek terkait lainnya.
5.	Joni Salminen, Soon-gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Waleed Akhtar, Bernard J. Jansen	Using Cipherbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot	2024	Cipherbot merupakan chatbot pendidikan yang menggunakan model bahasa besar untuk menjawab pertanyaan siswa mengenai materi pembelajaran yang diunggah oleh pendidik, telah diuji coba di ruang kelas. Cipherbot mampu menjawab 82,5% pertanyaan siswa,

				menunjukkan kemampuan yang terukur untuk menjawab pertanyaan siswa dengan beberapa ruang untuk perbaikan.
6.	Hansi K Radhakrishnan, N. G. J. Dias	Use of a ChatBot-Based Advising System for the Higher-Education System	2023	Penelitian ini memperkenalkan sebuah mekanisme yang layak dari sistem penasihat berbasis chatbot untuk menjembatani kesenjangan yang teridentifikasi antara kebutuhan pelajar dan ketersediaan sumber daya dengan mengotomatiskan proses pemberian nasihat di luar metode yang tersedia secara tradisional. Penelitian ini menggunakan model bahasa open source Large Language Model (LLM) yang dikombinasikan dengan basis pengetahuan khusus untuk mengatasi aspek-aspek penting yang dibutuhkan. Sistem ini menunjukan akurasi yang tinggi yaitu 89%
7.	Sumit Pandey, Srishti Sharma	A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning	2023	Penelitian ini mengembangkan dua chatbot berbasis retrieval dan berbasis generatif, masing-masing dengan enam desain. Di antara chatbot berbasis retrival, Vanilla Recurrent Neural Network (RNN) memiliki akurasi 83,22%, Long Short Term Memory (LSTM) memiliki akurasi 89,87%, Bidirectional LSTM (Bi-LSTM) memiliki akurasi 91,57%, Gated Recurrent Unit (GRU) memiliki akurasi 65,57% akurat, dan Convolution Neural Network (CNN) adalah 82,33% akurat. Sebagai perbandingan, chatbot berbasis generatif memiliki desain encoder-decoder yang 94,45% akurat

8.	Joni Salminen, Soon-gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Waleed Akhtar, dan Bernard J. Jansen	Using Cipherbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot	2024	Penelitian ini menggunakan model GPT-3.5 Turbo dengan teknik Retrieval-Augmented Generation (RAG) untuk mengembangkan chatbot edukasi bernama Cipherbot. Dataset terdiri dari materi pembelajaran dalam format PDF. Cipherbot mencapai tingkat keberhasilan 82.5% dalam menjawab pertanyaan mahasiswa, menunjukkan potensinya sebagai alat pendukung pembelajaran meskipun membutuhkan peningkatan lebih lanjut.
9.	Anggun Tri Utami Br. Lubis, Nazruddin Safaat Harahap, Surya Agustian, Muhammad Irsyad, dan Iis Afrianty	Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan)	2024	Penelitian ini menggunakan model Large Language Models (LLM) yang diterapkan melalui framework Langchain untuk mengembangkan sistem Question Answering (QAS) yang terintegrasi dengan chatbot Telegram. Dataset berasal dari dokumen UU No. 17 Tahun 2023 tentang Kesehatan, yang diproses menggunakan teknik chunking dan embeddings untuk membentuk basis pengetahuan. Pengujian dilakukan menggunakan metrik BERTScore dan ROUGE, dengan hasil rata-rata precision, recall, dan f1-score masing-masing sebesar 76%, 80%, dan 78%. Chatbot ini mampu memberikan jawaban yang relevan, meskipun masih terdapat keterbatasan dalam menangani pertanyaan di luar cakupan data yang diberikan.

2.10 Perbedaan Penelitian

Penelitian ini memiliki perbedaan dengan penelitian yang telah dilakukan sebelumnya, dimana pada penelitian (Galstyan & Martirosyan, 2024) memanfaatkan kemampuan model pembuatan teks untuk menghadirkan chatbot penasihat akademik yang bertujuan untuk mengembangkan sumber daya menyeluruh yang menggabungkan pengetahuan tentang universitas program studi untuk mahasiswa. Sedangkan penelitian ini, untuk mengembangkan chatbot virtual assistant untuk memudahkan pencarian informasi mengenai Universitas Sumatera Utara.

Kemudian perbedaan selanjutnya, dimana pada penelitian yang dilakukan oleh (Rachmat & Kesuma, 2024) dilakukan untuk memahami bagaimana implementasi API key Gemini dan apakah chatbot yang dibangun menggunakan Gemini dapat berfungsi sesuai dengan fungsionalitas Gemini LLM itu sendiri. Sedangan pada penelitian ini menerapkan teknologi Graph Based Retrieval-Augmented Generation (GraphRAG) dengan dukungan LangChain dan GPT.

Perbedaan selanjutnya, dimana pada penelitian yang dilakukan oleh (Odede Ingo Frommholz JAOdede, 2024) menyajikan JayBot yaitu system chatbot berbasis LLM yang bertujuan untuk meningkatkan pengalaman pengguna dosen, staf, calon mahasiswa dan mahasiswa di sebuah universitas di inggris. Sedangkan pada penelitian ini mengembangakan chatbot berbasis LLM menggunakan GraphRAG untuk memudahkan pencarian informasi mengenai USU dan membantu tenaga kependidikan dalam memberikan informasi secara akurat dan cepat.

BAB III

ANALISIS DAN PERANCANGAN SISTEM

3.1 Sumber Data

Dataset yang digunakan penulis pada penelitian ini dikumpulkan melalui metode web scraping yang dilakukan terhadap seluruh website Universitas Sumatera Utara (USU) serta subdomain-subdomain yang berhubungan dengan kemahasiswaan. Proses ini dilakukan untuk memperoleh data yang komprehensif mengenai konten, struktur, dan informasi yang relevan dengan layanan dan kegiatan kemahasiswaan. Pengumpulan data dilakukan dengan memanfaatkan teknik web scraping untuk mengekstrak informasi secara otomatis dari website USU dan subdomain yang terkait, seperti portal mahasiswa, UKM, biro kemahasiswaan, dan layanan lainnya. Metode ini memungkinkan peneliti mengumpulkan data secara menyeluruh dan efisien dari sumber-sumber publik yang tersedia. Hasil pengumpulan data ini selanjutnya digunakan sebagai dasar untuk mengimplementasikan GraphRAG pada Large Language Model (LLM) dalam pengembangan virtual assistant USU. Pendekatan ini diharapkan dapat meningkatkan kinerja sistem dalam memberikan layanan informasi kemahasiswaan yang lebih responsif dan akurat.

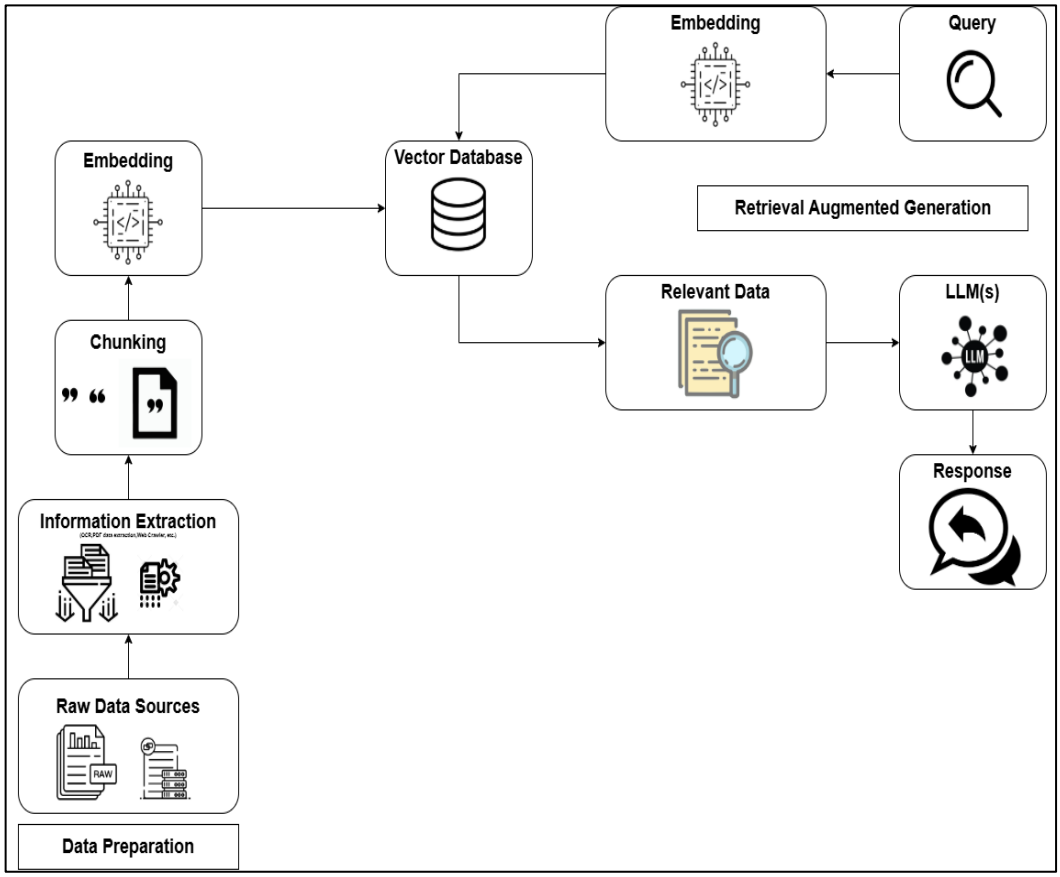
Tabel 3.1 Contoh Pertanyaan User

No	Contoh Pertanyaan User	Kategori
1	Saya ingin melihat kalender akademik USU, dimana saya bisa menemukannya?	Jadwal & Kegiatan
2	Bagaimana cara mendaftar UKM di USU?	Pendaftaran
3	Bagaimana saya mendapatkan informasi lengkap mengenai beasiswa dan bantuan keuangan untuk mahasiswa?	Informasi Akademik
4	Bagaimana proses pengajuan cuti akademik atau izin kuliah di USU?	Administrasi Akademik
5	Bagaimana prosedur pengajuan surat keterangan aktif kuliah di USU?	Administrasi Akademik

6	Saya butuh informasi terkait tata tertib kampus dan peraturan kemahasiswaan. Bagaimana saya dapat mengaksesnya?	Regulasi
7	Bagaimana prosedur pendaftaran ulang dan administrasi bagi mahasiswa baru?	Administrasi Akademik
8	Di mana saya bisa mendapatkan informasi lengkap tentang UKM dan kegiatan kemahasiswaan untuk maba?	Kegiatan Kemahasiswaan
9	Apa saja fakultas dan program studi yang tersedia di USU?	Informasi Akademik

3.2 Arsitektur Umum

Arsitektur umum untuk perancangan chatbot virtual assistant Universitas Sumatera Utara dapat dilihat pada gambar 3.2 berikut ini.



Gambar 3.1 Arsitektur Umum

3.1 Raw Data Source

Data mentah dikumpulkan dari berbagai sumber, seperti dokumen teks, PDF, database, atau file lainnya. Data ini biasanya tidak terstruktur atau semi-terstruktur.

3.2 Data Preparation

Data mentah dibersihkan dan diubah ke format yang lebih terstruktur, seperti menghapus duplikasi, mengisi data yang hilang, atau mengatur ulang formatnya agar siap untuk diproses lebih lanjut.

3.3 Information Extraction

Informasi penting diambil dari data yang telah dipersiapkan menggunakan teknik seperti OCR, Natural Language Processing (NLP), atau ekstraksi pola tertentu. Contoh hasilnya adalah entitas seperti nama, tanggal, atau angka.

3.4 Chunking

Data yang diekstraksi dipecah menjadi unit kecil (chunks) untuk memudahkan pengolahan. Potongan-potongan ini dapat berupa kalimat, paragraf, atau segmen teks.

3.5 Embedding

Setiap potongan data diubah menjadi representasi numerik atau vektor menggunakan algoritma embedding. Representasi ini memungkinkan komputer memahami hubungan antar data dalam bentuk matematika.

3.6 Vector Database

Vektor hasil embedding disimpan dalam basis data khusus untuk mendukung pencarian cepat berdasarkan kesamaan atau jarak antar vektor.

3.7 Query

Pengguna memberikan input berupa pertanyaan atau kata kunci yang ingin dicari informasinya dalam sistem.

3.8 Embedding (Query)

Pertanyaan atau query pengguna diubah menjadi vektor embedding menggunakan metode yang sama seperti pada data, sehingga dapat dibandingkan dengan data yang ada.

3.9 Retrival

Sistem membandingkan vektor query dengan vektor data dalam database untuk menemukan data yang paling relevan. Proses ini dilakukan dengan mengukur kesamaan antar vektor.

3.10 Relevant Data

Data yang dianggap paling relevan dengan query pengguna diambil dari database untuk diproses lebih lanjut.

3.11 LLM(s)

Data relevan yang telah ditemukan diproses oleh model bahasa besar (Large Language Model), seperti GPT, untuk menghasilkan jawaban atau respons yang sesuai dengan konteks query.

3.12 Response

Jawaban atau informasi yang telah diolah oleh model bahasa besar diberikan kembali kepada pengguna dalam format yang mudah dipahami.

DAFTAR PUSTAKA

’Wicaksana, M. P., ’Rahardandi, P. G., & Fauzan, M. (2024). Analisis Penerapan Chatbot: Survei. *Innovative: Journal Of Social Science Research*, *4*(4).

Arora, D., Singh, H. G., & Mausam. (2023). Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/2023.emnlp-main.468

Bariyah, S. H., & Imania, K. A. N. (2022). Pengembangan Virtual Assistant Chatbot Berbasis Whatsapp Pada Pusat Layanan Informasi Mahasiswa Institut Pendidikan Indonesia - Garut. *JURNAL PETIK*, *8*(1). https://doi.org/10.31980/jpetik.v8i1.1575

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. ., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). *Language models are few-shot learners. Advances in neural information processing systems*.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3). https://doi.org/10.1145/3641289

Galstyan, L. &, & Martirosyan, H. (2024). *SmartAdvisor-University-Chatbot*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. http://arxiv.org/abs/2407.21783

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3). https://doi.org/10.1007/s11042-022-13428-4

Odede Ingo Frommholz JAOdede, J. (2024). *JayBot – Aiding University Students and Admission with an LLM-based Chatbot*. https://github.com/Greenconsult/Jaybot-Chiir-Paper,

Quidwai, M. A., & Lagana, A. (2024). *A RAG Chatbot for Precision Medicine of Multiple Myeloma*. https://doi.org/10.1101/2024.03.14.24304293

Rachmat, N., & Kesuma, D. P. (2024). *Implementasi Large Language Models Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android* (Vol. 4, Issue 1). https://journal.umgo.ac.id/index.php/juik/index

Radhakrishnan, H. K., & Dias, N. G. J. (2023). *Use of a ChatBot-Based Advising System for the Higher-Education System*.

Rantung, P. . (2023). *TEKNIK-TEKNIK PEMROSESAN BAHASA ALAMI (NLP)*. Lakeisha.

Shafee, S., Bessani, A., & Ferreira, P. M. (2024). *Evaluation of LLM Chatbots for OSINT-based Cyber Threat Awareness*. http://arxiv.org/abs/2401.15127

Khan, M. S. (2018). Impact of Information Technology on Education Sector. International Journal of Advanced Research in Computer Science and Management Studies, 6(5), 1-5.

Haristiani, N., 2019, November. Artificial Intelligence (AI) chatbot as language learning medium: An inquiry. In Journal of Physics: Conference Series (Vol. 1387, No. 1, p. 012020). IOP Publishing.

Hussain, W., Shad, A. B., & Jabeen, F. (2021). Chatbots in Higher Education: A Review. International Journal of Educational Technology in Higher Education, 18(1), 1-15.

M. U. Hadi, Q. Al-Tashi, R. Qureshi dan e. a. , “Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects,” TechRxiv, vol. 4, Seoptember 2023, https://doi.org/10.36227/techrxiv.23589741.v3.

Hu, Krystal, “ChatGPT sets record for fastest-growing user base - analyst note,”. Internet: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, 2 Februari 2023 [5 Januari 2024].

Fatima, S., Aslam, N., & Hussain, M. (2020). Evaluating the Performance of Chatbots in Education: A Systematic Review. Journal of Educational Computing Research, 58(5), 1007-1027.