BAB I

PENDAHULUAN

1.1 Latar Belakang

Kehadiran teknologi informasi dalam pelayanan sumber informasi akademik menjadi aspek yang sangat penting di perguruan tinggi. Pelayanan yang diberikan oleh perguruan tinggi menjadi cerminan kualitas lembaga tersebut (Ajiz dkk., 2023). Universitas Pendidikan Ganesha (Undiksha) memiliki berbagai sumber informasi akademik yang penting bagi civitas akademikanya. Sumber akademik tersebut diantaranya pedoman akademik, kalender akademik dan lain sebagainya. Informasi ini merupakan komponen utama dalam mendukung proses belajar mengajar, penelitian, serta kegiatan akademik lainnya.

Baik kalender akademik akademik maupun pedoman akademik di Undiksha umumnya memiliki jumlah halaman yang cukup banyak. Hal tersebut menyebabkan pencarian informasi secara manual menjadi kurang efisien. Pengguna sering kali harus membuka halaman demi halaman untuk menemukan informasi spesifik yang mereka butuhkan. Berdasarkan survei yang dilakukan peneliti terhadap lebih dari 300 mahasiswa Undiksha, menunjukan bahwa 73% mahasiswa cukup kesulitan dalam mencari informasi akademik dengan waktu rata-rata 5-10 menit. Hal tersebut mengindikasikan ketidakefisienan dalam mencari informasi akademik.

Teknologi chatbot menjadi salah satu inovasi dalam menyediakan layanan informasi akademik yang responsif dan efisien (Ajiz dkk., 2023). Seiring dengan berkembangnya teknologi chatbot, saat ini mulai muncul chatbot berbasis Model

Bahasa Besar (*Large Languange Model* atau LLM) dalam menghasilkan informasi berbasis teks. Model Bahasa Besar (*Large Language Model* atau LLM) adalah model *Machine Learning* yang dilatih dengan kumpulan data teks dalam jumlah yang banyak. Menurut (Chang dkk., 2024) peranan LLM saat ini semakin populer dalam bidang akademis, bidang industri maupun untuk keperluan sehari-hari. Beberapa LLM yang kompeten dalam pemrosesan bahasa alami (NLP) diantaranya yaitu BERT, GPT (OpenAI), LlaMA-2 (Facebook) dan Claude (AnthropicAI) (Zhao dkk., 2024).

Menurut Yao dkk. (2023), meskipun tampaknya LLM seperti GPT dan LlaMa memiliki pengetahuan yang luas dan adaptif terhadap banyak tugas, kita sebagai pengguna belum dapat sepenuhnya percaya terhadap jawaban yang diberikan oleh LLM. Hal ini dikarenakan LLM dapat mengalami "halusinasi" memberikan jawaban yang salah dan tidak mendasar (Farquhar dkk., 2024). Dataset LLM cenderung berasal dari informasi publik dan terbatas untuk mengakses data-data pribadi seperti data internal kampus atau instansi lain yang tidak terpublikasi.

Untuk mengatasi halusinasi pada LLM, salah satu teknologi yang dapat menjadi solusi adalah *Retrival-Augmented Generation* (RAG). Dikutip dari (Wu dkk., 2024), *Retrieval-Augmented Generation* (RAG) merupakan sebuah pendekatan inovatif dalam pengembangan model bahasa besar (LLMs) yang bertujuan untuk mengatasi beberapa keterbatasan utama LLMs konvensional. Pada RAG, data yang berbasis teks dipecah menjadi banyak bagian (*chunk*) (Wu dkk., 2024). *Chunk* tersebut kemudian direpresentasikan menjadi vektor dengan menggunakan model *embedder* dan kemudian disimpan dalam basis data vektor. Ketika pengguna bertanya, *query* pengguna akan ubah menjadi representasi vektor dan dicari

kemiripan pada basis data vektor (*Retrival*). Setelah dokumen yang sesuai didapatkan maka dilakukan penambahan *prompt* pada dokumen yang didapatkan agar lebih dipahami oleh LLM (*Augmented*). Terakhir adalah proses mehasilkan teks alami berdasarkan gabungan antara *prompt* dan dokumen dengan menggunakan LLM (*Generation*) .

Dalam pengembangannya, sistem RAG konvensional juga memiliki kelemahan dimana proses *retrieval* juga seringkali memberikan hasil yang tidak akurat (Yan dkk., 2024). Hal ini dikarenakan *query* pengguna dieksekusi secara langsung tanpa melalui proses pembersihan *query*. Untuk mengatasi permasalah tersebut, kini telah dikembangkan konsep *Advance-RAG. Advance-RAG* memodifikasi proses RAG konvensional yang awalnya hanya fokus pada proses *retrieval* dan *generation,* kemudian dikembangkan menjadi *pre-retrieval*, *retrieval*, *post-retrieval* dan *generation* (Yuan dkk., 2024). *Pre-retrieval* bertujuan untuk memanipulasi *query* pengguna sehingga menjadi lebih jelas dan meningkatkan relevansi pada tahap *retrieval* dengan mengikutsertakan LLM didalamnya. Sedangkan pada *post-retrieval* data dipilah berdasarkan kesesuaian dengan kueri untuk menghasilkan respons yang lebih akurat. Ini membantu mengurangi masalah halusinasi dan memungkinkan model untuk menghasilkan jawaban yang lebih akurat.

Menurut Jeong (2024), *Advance RAG* terbagi menjadi tiga kategori yaitu *Self-RAG*, *Corrective-RAG* dan *Adaptive-RAG*. *Self-RAG* berfokus pada memberikan jawaban diawal, kemudian mengekstrak jawaban tersebut menjadi konteks pertanyaan. Sedangkan *Corrective RAG* bekerja dengan mengoreksi hasil pencarian dengan menghilangkan informasi tidak relevan. Terakhir yaitu *Adaptive-*

RAG yakni menyesuaikan dengan pertanyaan pengguna untuk mengarahkan ke jenis pencarian yang sesuai.

Salah satu inovasi yang menarik dalam mengembangkan sistem berbasis RAG adalah dengan menggunakan arsitektur *multi-agent.* Dikutip dari Guo dkk. (2024), *agent* dalam arsitektur *multi-agent* adalah sebuah fungsi otonom yang dapat berupa LLM, *retriever* ataupun fungsi pemrograman biasa. Arsitektur *multi-agent* dimaksudkan untuk membangun agen-agen otonom yang memiliki tugas dan kemampuan berbeda. Misalnya, satu *agent* dapat berfokus pada pemahaman kontekstual, sementara *agent* lain dapat berfokus pada pengambilan informasi yang relevan dari basis data. Setiap agen dapat berinteraksi dan saling bertukar data untuk menghasilkan jawaban yang informatif dan faktual.

Arsitektur *Multi-Agent* dan RAG menjadi solusi untuk mengembangkan sistem informasi yang efisien dan akurat. Dengan menggabungkan LLM, RAG dan agen-agen khusus untuk tugas tertentu memungkinan untuk memanajemen tugas, interaksi dengan bahasa alami serta koordinasi antar agen (Xu dkk., 2024). Untuk mengurangi kesulitan mahasiswa dalam me, sistem berbasis chatbot akademik "AKASHA" dengan pendekatan *Corrective Retrieval Augmented Generation* (CRAG) menjadi pilihan yang tepat dan efisien. Dengan adanya *chatbot* ini diharapkan dapat menjadi menjadi media pencarian informasi akademik yang cepat dan efisien bagi civitas akademika.

1.2 Rumusan Masalah

1. Bagaimana membangun Chatbot *Generative AI* "AKASHA" dengan menggunakan pendekatan *Multi-Agent Corrective-RAG* untuk informasi akademik Undiksha dengan dengan antarmuka aplikasi web?
2. Bagaimana mengevaluasi kinerja Chatbot *Generative AI* "AKASHA" dengan menggunakan pendekatan *Multi-Agent Corrective-RAG* untuk informasi akademik Undiksha?

1.3 Tujuan

1. Membangun Chatbot *Generative AI* "AKASHA" dengan menggunakan pendekatan *Multi-Agent Corrective-RAG* untuk informasi akademik Undiksha dengan dengan antarmuka aplikasi web.
2. Mengevaluasi kinerja Chatbot *Generative AI* "AKASHA" dengan menggunakan pendekatan *Multi-Agent Corrective-RAG* untuk informasi akademik Undiksha.

1.4 Batasan Masalah Penelitian

1. Penelitian ini hanya menggunakan sumber data dari pedoman akademik Undiksha, kalender akademik Undiksha terbaru dan sumber data akademik lain relevan.
2. Pengembangan aplikasi ini terbatas pada penggunaan *Corrective Retrieval Augmented Generation* (CRAG) dengan menggunakan *framework* LangChain dan LangGraph. Penggunaan metode dan atau *framework* lain tidak termasuk dalam lingkup penelitian ini.

3. Pengembangan chatbot ini hanya menggunakan model-model yang dikembangkan oleh OpenAI dan tidak menggunakan model bahasa besar yang dikembangkan oleh vendor lain.
4. Penelitian ini hanya berfokus pada pengembangan model *Corrective Retrieval Augmented Generation* (CRAG) dengan *platform* web sebagai media untuk menguji fungsionalitas model. Proses pengembangan web bukan menjadi fokus utama pada penelitian ini dan web hanya ditujukan untuk visualisasi pembuatan basis pengetahuan RAG dan tanya jawab pada model RAG.
5. Metode *chunking* hanya menggunakan fungsi RecursiveCharacterTextSplitter dari LangChain dengan konfigurasi *chunk size* sebesar 800 karakter dan *chunk overlap* sebesar 200 karakter.

1.5 Manfaat Penelitian

1. Manfaat Teoritis
   - Menyediakan kontribusi ilmiah dalam pengembangan teknologi *Retrieval-Augmented Generation* (RAG), khususnya *Corrective-RAG* dengan kombinasi sistem *Multi-agent* untuk meningkatkan akurasi pencarian dan pengelolaan informasi.
   - Menambah referensi dalam literatur terkait implementasi Model Bahasa Besar *(Large Language Model*/LLM) pada aplikasi pencarian informasi di bidang akademik.
   - Mengembangkan model konseptual dan teknik *pre-retrieval*, *retrieval*, *post-retrieval*, dan *generation* yang dapat diadaptasi dalam penelitian-penelitian berikutnya.

2. Manfaat Praktis

- Mempermudah mahasiswa dan staf akademik Universitas Pendidikan Ganesha (Undiksha) dalam mengakses informasi akademik seperti pedoman akademik, kalender akademik, dan dokumen terkait secara cepat dan tepat.
- Mengurangi waktu dan kesulitan dalam pencarian informasi akademik, sehingga meningkatkan efisiensi layanan informasi di lingkungan kampus
- Memberikan solusi berbasis teknologi yang praktis dan inovatif dalam mendigitalisasi layanan informasi akademik Undiksha.