AI Chatbot Development for Computer Science Curriculum Based Academic Consultation using Retrieval-Augmented Generation (RAG) = Pengembangan Chatbot AI untuk Bimbingan Akademik Berbasis Kurikulum Ilmu Komputer dengan Retrieval-Augmented Generation (RAG)

Daniel Christian Mandolang, author

Abstrak

Academic consultation is a crucial part of students’ university journey. As the curriculum evolves, there is a growing need for a chatbot application to assist in this process. This development aims to enhance the effectiveness and efficiency of academic consultations. However, due to the nature of large language models (LLMs), which may generate ”hallucinations,” and the application of Retrieval-Augmented Generation (RAG) techniques in chatbot development, our chatbot will adopt this approach. In RAG-based chatbot development, any corpus can be used to obtain context for generation. In this research, the corpora utilized include a vector database and a knowledge graph database. This study focuses on improving performance in terms of answer relevance to questions within the academic consultation context, while also considering response time. Three architectures are tested in our chatbot, such as Hierarchical Tree Retrieval and Graph-Augmented Retrieval which is a vector-based retrieval, as well as Knowledge Graph Retrieval, to answer prerequisite-related questions that are particularly challenging for Collapsed Vector Retrieval. The final chatbot model is integrated into an application deployed on-premises at the Center for Computer Science (Pusilkom) of the Faculty of Computer Science, University of Indonesia. The best retrieval architecture in vector-based retrieval is the Graph-Augmented Retrieval which combines the advantage of vector-based chatbot and the ability to recognize prerequisite information, shown by the relatively high mAP of 0.744 and hit rate of 0.904 that outperforms other vector architectures. Additionally, the Knowledge Graph Retrieval improved the answer relevance significantly in answering prerequisite-related questions, shown by 100% answer accuracy improvement in prerequisite-related questions compared to the base model. Finally, among the models tested, Knowledge Graph Retrieval achieves the highest answer accuracy. However, the Graph-Augmented Retrieval model is considered to be the best for an on-premises chatbot, providing 76% answer accuracy and an average response time of 8.49 seconds. This represents a 6.5 seconds reduction in the response time compared to the Knowledge Graph Retrieval model. All of our chatbot models have been successfully deployed on the specified server and are utilizing all GPU resources optimally. However, the chatbot application is not able to provide an acceptable user experience utilizing the computing resource in the specified server when there are multiple concurrent users (more than one).

.....Konsultasi akademik merupakan bagian penting dari perjalanan universitas mahasiswa. Seiring dengan evolusi kurikulum, ada kebutuhan yang semakin besar untuk aplikasi chatbot yang dapat membantu dalam proses ini. Pengembangan ini bertujuan untuk meningkatkan efektivitas dan efisiensi konsultasi akademik. Namun, karena sifat dari large language models (LLMs), yang dapat menghasilkan "hallucinations," dan penerapan teknik Retrieval-Augmented Generation (RAG) dalam pengembangan chatbot, chatbot kami akan mengadopsi pendekatan ini. Dalam pengembangan chatbot berbasis RAG, corpus apa pun dapat digunakan untuk memperoleh konteks untuk generasi. Dalam penelitian ini, corpus yang digunakan meliputi database

vektor dan database knowledge graph. Studi ini berfokus pada peningkatan kinerja dalam hal relevansi jawaban terhadap pertanyaan dalam konteks konsultasi akademik, sambil juga mempertimbangkan waktu respons. Tiga arsitektur diuji dalam chatbot kami, seperti Hierarchical Tree Retrieval dan Graph-Augmented Retrieval yang merupakan retrieval berbasis vektor, serta Knowledge Graph Retrieval, untuk menjawab pertanyaan terkait prasyarat yang sangat menantang bagi Collapsed Vector Retrieval. Model chatbot akhir diintegrasikan ke dalam aplikasi yang diterapkan di server lokal di Center for Computer Science (Pusilkom) Fakultas Ilmu Komputer, Universitas Indonesia. Arsitektur retrieval terbaik dalam retrieval berbasis vektor adalah Graph-Augmented Retrieval yang menggabungkan keuntungan chatbot berbasis vektor dan kemampuan untuk mengenali informasi prasyarat, yang ditunjukkan oleh mAP yang relatif tinggi yaitu 0,744 dan hit rate 0,904 yang mengungguli arsitektur vektor lainnya. Selain itu, Knowledge Graph Retrieval meningkatkan relevansi jawaban secara signifikan dalam menjawab pertanyaan terkait prasyarat, yang ditunjukkan dengan peningkatan akurasi jawaban sebesar 100% pada pertanyaan terkait prasyarat dibandingkan dengan model dasar. Akhirnya, di antara model yang diuji, Knowledge Graph Retrieval mencapai akurasi jawaban tertinggi. Namun, model Graph-Augmented Retrieval dianggap sebagai yang terbaik untuk chatbot lokal, dengan memberikan akurasi jawaban sebesar 76% dan waktu respons rata-rata 8,49 detik. Ini menunjukkan pengurangan waktu respons sebesar 6,5 detik dibandingkan dengan model Knowledge Graph Retrieval. Semua model chatbot kami telah berhasil diterapkan di server yang ditentukan dan memanfaatkan semua sumber daya GPU secara optimal. Namun, aplikasi chatbot ini tidak dapat memberikan pengalaman pengguna yang dapat diterima dengan memanfaatkan sumber daya komputasi di server yang ditentukan ketika ada banyak pengguna yang berjalan bersamaan (lebih dari satu).