Perfect Studymate

Juyong Rhee[1], Yewon Chun[2], Jihee Hwang[3], Jorge Alcorta[4],

[1] Department of Chinese Language and Literature, SungKyunKwan University
qpsvs0131@gmail.com
[2] Department of Statistics, SungKyunKwan University 1000daughther@g.skku.edu
[3] Department of Korean Language and Literature, SungKyunKwan University
jihee336723@gmail.com
[4] Department of Software, SungKyunKwan University 2024319370@g.skku.edu

Abstract. This paper proposes a Retrieval-Augmented Generation (RAG)-based AI chatbot to address the limitations of existing AI tools like ChatGPT in academic settings. While ChatGPT often provides general and inaccurate responses, the RAG-based chatbot will retrieve precise, course-specific information from uploaded materials, improving learning efficiency for computer science students. The system will leverage technologies such as LangChain, ChromaDB, and LLaMa to build a personalized study tool. By offering tailored responses based on course documents, this solution aims to lower the learning curve and provide a more effective educational experience.

Keywords: LangChain · RAG · LLM · Chatbot

1 Introduction

As AI becomes increasingly integrated into daily life, tools like ChatGPT are frequently utilized by students, especially those in computer science, to enhance their learning efficiency. While widely adopted, ChatGPT presents several limitations as an educational tool. First, the information provided by ChatGPT often differs from course-specific materials, particularly in terminology, definitions, and use cases, as its training data is broad and lacks domain-specific depth. Second, ChatGPT's responses are influenced by prior conversational context, which can introduce inaccuracies. Lastly, while ChatGPT excels at delivering general information, it is not designed to optimize the learning process for specific courses. Thus, relying on it to reduce the learning curve may have unintended consequences.

In contrast, we propose a Retrieval-Augmented Generation (RAG)-based AI chatbot tailored for computer science students. This system will allow learners to engage with course materials directly by retrieving content from uploaded documents, ensuring that responses are aligned with the curriculum. Unlike ChatGPT, which may generate plausible but inaccurate information, the RAG-based chatbot will strictly depend on the provided materials. This approach is expected to lower the learning curve by offering precise, course-specific answers, thereby creating a more effective and enjoyable learning experience for computer science education.

For the system, we will first gather and prepare the necessary course materials, which will serve as the knowledge base for the chatbot. Next, we will implement the RAG model, ensuring it can retrieve the correct information from these materials. At the same time, backend and frontend will also be in development process to connect with the chatbot. Finally, the system will eventually be integrated and tested to confirm it provides accurate and helpful answers based on the uploaded documents.

2 Motivation and Objective

2.1 Motivation

In today's fast-paced academic environment, with the increasing adoption of digital education tools such as digital textbooks, students are often overwhelmed by the sheer volume of digital lecture materials they need to manage. As the amount of content increases, the challenge of efficiently finding and utilizing critical information has become a common struggle. We have personally experienced the frustration of searching through multiple PDFs, trying to locate a single term or piece of information. Tools like GoodNotes are widely used, yet they fall short of offering a comprehensive search across all course materials, which has only heightened this challenge.

Moreover, while general AI tools such as ChatGPT are helpful in answering questions, they often fail to address the specific needs of students in the context of their courses. We have noticed that generic AI tools sometimes provide inaccurate or irrelevant answers because they aren't designed with academic materials in mind. This lack of precision is frustrating, especially when students are seeking exact information from their course lectures.

Additionally, the cost of uploading materials to existing AI services is another barrier that prevents students from fully leveraging these tools for their academic needs. Even when lecture materials are uploaded, they are referenced only temporarily, limiting the ability to create a personalized and long-term study assistant.

Through these experiences, we recognized a strong need for a customized solution—one that allows students to effectively search, study, and quiz themselves based on their own lecture materials, providing a much-needed companion for modern academic success.

2.2 Objective

The primary objective of the Perfect Studymate project is to develop a customized chatbot that assists students in effectively managing and utilizing their course materials. The system will allow students to upload their lecture materials (e.g., PDFs), which will be processed using Retrieval-Augmented Generation (RAG). By leveraging the OpenAI API, the chatbot will provide services below.

- Provide accurate answers to course-related questions based on the uploaded materials.
- Help students locate specific terminology or concepts within their course materials.
- Generate customized quizzes derived from the course content to aid in study preparation.
- Summarize uploaded resources using proper terminologies.

This project aims to enhance the overall learning experience by providing a personalized and efficient study tool for students.

3 Background and Related Work

3.1 Background

LangChain LangChain is an open source SDK for building applications easily based on large language models (LLMs). LangChain provides tools and abstractions to improve the customization, accuracy, and relevancy of the information the models generate. Below is a brief description of Langchain flow.

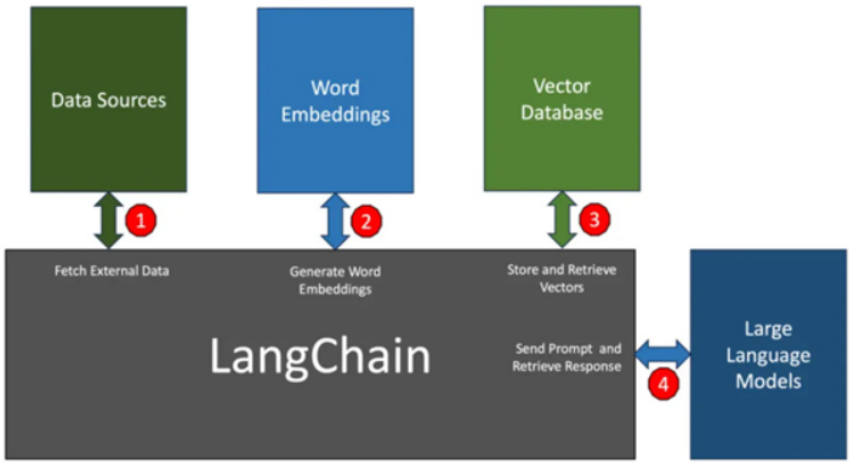

Fig. 1. flowchart of Langchain

Firstly, LangChain includes components that allow LLMs to access new data sets without retraining. In some cases, applications need to retrieve data from external sources such as PDFs, web pages, CSVs, relational databases to build the context for LLM. Langchain integrates seamlessly with modules that can access and retrieve data from different sources.

Secondly, the data retrieved from some external sources must be transformed into vectors. Only then will the text be passed to the word embedding model related to LLM. Langchain selects the optimal embedding model based on the selected LLM.

Subsequently, the generated embeddings are stored in a vector database for similarity retrieval. Langchain supports easy storage and retrieval of vectors from sources ranging from in-memory arrays to hosting vector databases.

Lastly, Langchain supports mainstream LLM provided by OpenAI, Cohere, and AI21 and open-source LLM provided by the Hugging Face. The list of supported models and API endpoints is growing rapidly.

ChromaDB chromaDB is a vector database. A vector database is a database that stores data in vector form beyond a general relational database. Vector databases are useful for storing, inquiring, and retrieving vector-embedded semi-structured, unstructured data. The values of the vector database are numerically

embedded, so it is easy to calculate the similarity of several stored values. This is why vector DB is used as a database for semi-structured and unstructured AI models.

ChromaDB has the advantage of making quick data inquiries and being simple to use. It ensures swift access to data without the latency associated with disk-based systems. This approach enhances the responsiveness of AI applications that demand real-time interactions with large datasets. Also, the simplicity of Chroma DB's API streamlines the development process by providing developers with a straightforward interface to interact with the database.

LLaMa LLaMa is a open-source large language model designed to help researchers advance their work in this subfield of AI. It can perform complex tasks related to text, such as text generation, conversation, and summary. llama has more strength in fine tuning than other language models. This means that the model's response can be more domain-specific.

3.2 Related Work

Sharly AI Summarizer Sharly AI summarizer tool is powered by artificial intelligence and provides concise summaries of any content, including PDF, general documents, articles, audio files, or presentations. Users upload their files and let the AI do the work. The AI summary generator reads through the file, identifies the main themes and the most essential details. summarizing into a thing of the past.

In addition to the summarization, this service provides functions such as quiz generation through an interactive chatbot, but there is a problem that they do not generate any response after user request model to generate quiz.

Quizlet Quizlet service provides users an AI-based learning solutions like generating dictionary and excercise based on textbook. Recently, it also launch Q-chat service which provides virtual AI tutor experience to users. But Actually this service is available on Austrailia, Canada, Germany and some other western countries except Asia and Africa. And in order to use this service, there is a limitation that you have to pay a certain amount of money by subscribing to Quizlet's service every month.

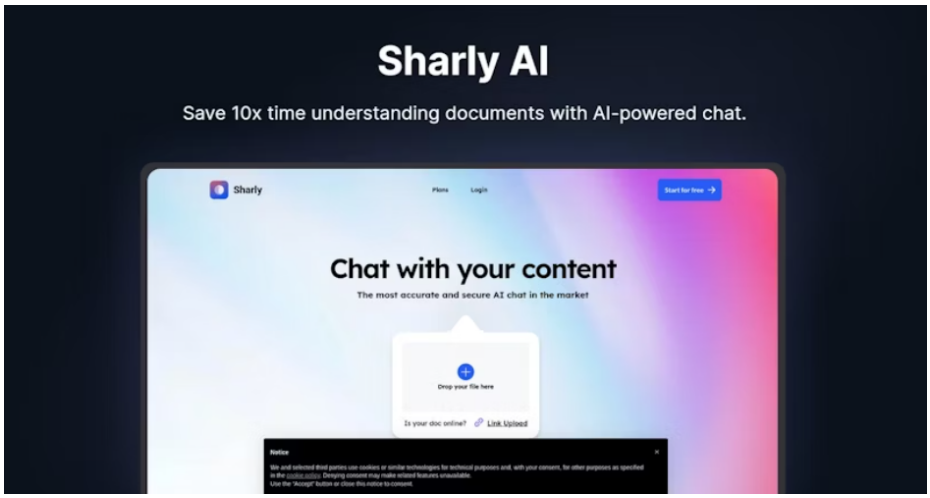

Fig. 2. screenshot of Sharly AI implementation page

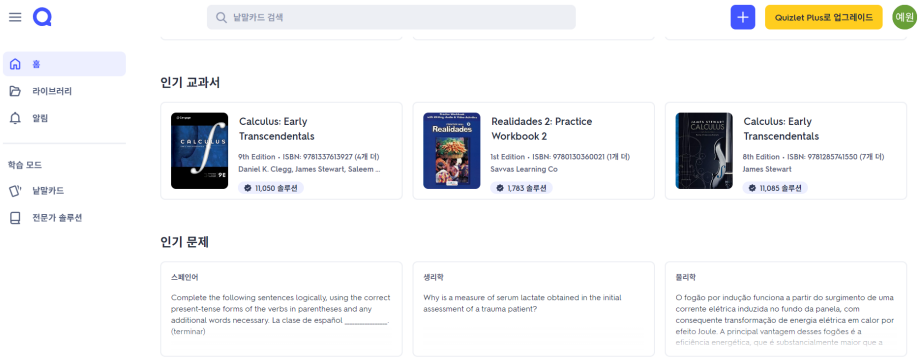

Fig. 3. screenshot of Quizlet. Quizlet is still unstable in terms of launching a new chatbot service because it originally have focused on textbook solution services.

4 Problem Statement and Proposed Solution

4.1 Problem Statement

Students face difficulties in retrieving precise and relevant information from course materials, often relying on generalized AI tools that introduce hallucinations or bias. This leads to confusion and inefficiency, as the answers provided may not be grounded in the specific course content. Based on appropriate resources, it is necessary to develop a model that can satisfy the needs of students who want to use static learning content more dynamically.

4.2 Proposed Solution

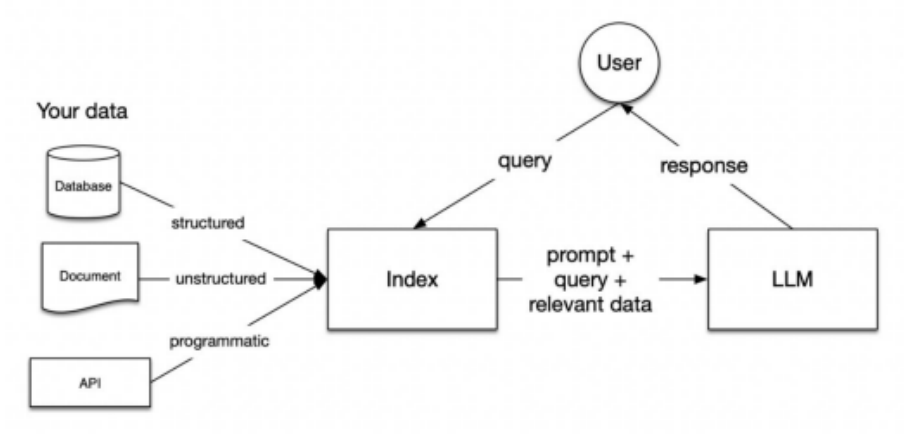

Fig. 4. How RAG-Based Applications Work

We propose a RAG-based AI chatbot system that allows students to upload course materials, with AI answering questions solely based on these materials. This solution will provide accurate, course-specific answers, preventing hallucinations and improving academic performance by ensuring that all responses align with the professor's material, alongside interactive features which will make learning a more enjoyable and overall positive experience.

5 Planning in Detail

5.1 Roles

Each team member primarily focuses on a specific area, ensuring expertise and efficiency, but we are committed to frequent communication and close cooperation throughout all tasks.

Name	**Roles**	**Description**
Jihee Hwang	FrontEnd Dev.	Responsible for designing and implementing the user interface,ensuring a user-friendly experience.
Juyong Rhee	BackEnd Dev.	Manages backend logic, API specifications, and integration, ensuring smooth data handling and functional service operations.
Jorge Alcorta	ML Modeling	Focuses on AI architecture design and development to incorporate intelligent features into our platform.
Yewon Chun	Documentation	In charge of continuously documenting the development process and finalizing project documentation, keeping track of all details and changes.

5.2 Schedule

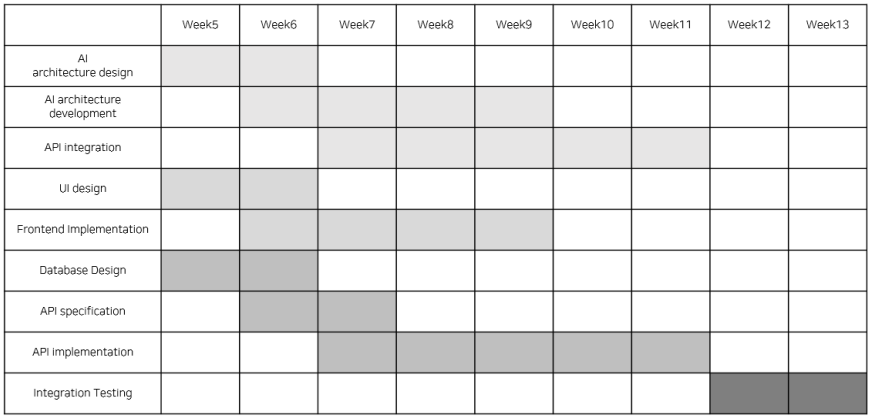

Fig. 5. this is brief schedule of ongoing development