ISSN: 2473-4901
v10 n6132

User Experiences in a RAG-Empowered Application

Shingo Kise
kiseshingo1@cityuniversity.edu

Sam Chung
chungsam@cityu.edu

MS in Computer Science (MSCS)
School of Technology & Computing (STC)
City University of Seattle (CityU)

Abstract

With the widespread use of Large Language Models (LLMs) in various applications, there has been growing interest in leveraging their capabilities to improve user experiences and streamline processes. However, given the availability of various LLMs and Retrieval-Augmented Generation (RAG) systems, it is crucial to understand the differences between these technologies to effectively implement them and maximize their potential for providing better services. This paper investigates whether a RAG-empowered mobile app can enhance user experiences by delivering more relevant responses to specific user inquiries than a system equipped solely with an LLM. We focus on RAG, facilitated by LangChain, and compare its effectiveness to that of LLMs. Students can benefit from internship placement systems guided by Natural Language Processing (NLP)-powered applications, improving educational outcomes and job prospects. We developed a chatbot-based internship placement system using React Native, integrating the ChatGPT API and LangChain for personalized, relevant responses. By evaluating chatbot responses using the RAG Assessment (Ragas) framework with metrics such as context precision, context recall, faithfulness, and answer relevancy to measure the quality of the RAG pipeline, we found that the RAG-empowered system consistently delivered more context-specific answers. Additionally, a qualitative comparison revealed that the LLM system tended to produce more generic responses compared to the RAG system. RAG systems can enhance the efficiency and effectiveness of internship placements by offering tailored assistance. Our findings highlight the potential of advanced NLP technologies to revolutionize applications such as chatbots, promoting innovation and enhancing user experiences.

Keywords: Large Language Model, LLM, Retrieval Augmented Generation, RAG, Case Study

1. INTRODUCTION

Natural language processing (NLP) has revolutionized in recent years due to the development of large language models (LLMs), which provide remarkable capacities for text generation, understanding, and manipulation. As LLMs have become more popular, cutting-edge methods like Retrieval Augmented Generation (RAG) have also surfaced, which facilitate offering substitute answers for a range of NLP problems.

Although LLMs and RAG share the goal of solving complex linguistic problems, their underlying architectures, approaches, and performance results are quite different. LLMs generate responses based solely on pre-trained data, while RAG systems combine retrieval from external sources with generation, allowing for more accurate, specific, and up-to-date responses (Zhao et al., 2024). We challenge whether a RAG-empowered mobile app can provide better user experiences by delivering more relevant responses to specific user inquiries compared to one equipped with only an LLM.

As a demo case study, we choose an internship matching program. Students at a university can

 ISSN: 2473-4901
v10 n6132

benefit from a system where they can gain valuable outside-of-class experiences, such as mentorship programs, internship matching programs, and more. Several case studies have been conducted on the development of web-based internship placement applications (Abdullah et al., 2017; Chaurasia, 2023). By applying an LLM or RAG to the system to guide each user based on their needs, we can significantly improve the quality of student's education experience and post-graduation job landing probabilities. This demonstration aims to showcase the potential of incorporating an LLM and an RAG into an internship placement system, instilling hope about the future of such systems.

The proposed approach entails developing a demonstration chatbot mobile application in React Native. This application seamlessly integrates with the OpenAI API to generate responses to user inquiries for the LLM system. Furthermore, LangChain is used within the application to provide answers to internship-related questions derived from files for the RAG system. Users can engage in interactive and informative interactions by combining these advanced technologies and receiving relevant responses to their queries in real-time. This novel combination of an LLM and RAG aims to improve the user experience by giving users a thorough understanding of internship opportunities and facilitating smoother interactions with the internship placement system, making them feel more engaged and informed.

2. BACKGROUND

Recognizing the importance of internship experiences and seeking to improve accessibility, there has been some interest in using technology to refine internship placement processes and provide students with more valuable industry opportunities. Several studies have been conducted to develop web-based internship placement systems that are specific to each university. For instance, Chaurasia (2023) developed an application using Python, the Django framework, and a MySQL database to enable fast and easy access to placement procedures and related activities. Similarly, Abdullah et al. (2017) used Adobe Dreamweaver CC 2014 and Cross-Platform, Apache, MariaDB, PHP, and Perl (XAMPP) to address manual management inefficiencies in their internship program, which was plagued by cumbersome paper-based processes and insufficient coordination among academics, industry stakeholders, and students. In the same way, others developed a recommendation system using technologies like C#, JavaScript, CSS, HTML, and MySQL. While these studies have successfully developed applications and received positive usability feedback from their test users, it is important to note that the primary goal of these systems has been to simplify the internship program process rather than to increase internship acquisition rates or overall employment outcomes.

React Native is an open-source framework developed by Facebook for building cross-platform mobile applications. React Native, which uses the popular React JavaScript library, allows developers to create mobile apps using tools and concepts familiar to web developers. One of its primary benefits is the ability to write code once and deploy it across multiple platforms, including iOS and Android, without compromising performance or user experience. React Native uses a declarative programming model, which allows developers to define UI components in JavaScript XML (JSX) syntax and then translate them into native User interface (UI) elements. This approach ensures that the resulting apps have a native appearance and feel, giving users a seamless experience (React Native RSS, n.d.).

LLMs are advanced AI models that have been trained on massive text datasets. These models, such as OpenAI's Generative Pre-trained Transformer (GPT) series, have transformed NLP tasks with their remarkable language understanding and generation capabilities. LLMs can comprehend and generate human-like text on a wide range of topics, making them extremely useful for a variety of NLP applications (Yang et al., 2024).

The OpenAI API, developed by OpenAI, gives developers access to cutting-edge LLMs, allowing them to use their powerful capabilities to build applications. The OpenAI API allows developers to integrate an LLM into their projects without having to train or fine-tune the models themselves. The API provides a variety of endpoints for tasks such as text generation, text completion, language translation, and more, allowing developers to easily incorporate advanced NLP capabilities into their applications (OpenAI platform, n.d).

RAG is an AI framework that enhances LLMs with accurate, current information from an external knowledge base (Martineau, 2024). LangChain is a framework that simplifies the development of RAG applications by integrating LLMs with external data sources. It offers resources for building modular chains that incorporate data

ISSN: 2473-4901
v10 n6132

from databases and APIs. This approach enhances the relevance and accuracy of the generated content by grounding it in external knowledge (LangChain, n.d.).

Ragas is a framework used to quantitatively evaluate the performance of RAG pipelines. It provides metrics and methods to assess the effectiveness of RAG systems by comparing generated responses with ground truth answers (Ragas, n.d.).

3. RELATED WORK

The rapid advancement of LLMs has revolutionized the field of NLP, providing unprecedented capabilities in addressing diverse NLP tasks and real-world applications (Yang et al., 2024). Integrating cutting-edge technologies such as OpenAI's GPT into real-world web and mobile applications has become more accessible in modern application development, providing practitioners with enhanced capabilities in NLP tasks (Odede & Frommholz, 2024). Furthermore, the development of RAG frameworks such as LangChain has made it easier to integrate LLMs into applications and helped improve LLM models. Recent studies by Gautam and Purwar (2024) have demonstrated that RAG systems can enhance the accuracy and relevance of responses, making them competitive with commercial LLMs. RAG enables LLMs to interact with various data sources and environments, allowing for model customization.

The application of LLMs and related frameworks in academia has led to practical innovations in the educational system. For instance, student assistant chatbots built on top of ChatGPT are a common application for LLM technology. These specialized chatbots, unlike traditional ones, focus on specific academic disciplines, providing tailored responses to student queries. The integration of OpenAI's advanced LLMs (GPT-3.5 turbo) with the LangChain framework and vector databases allows chatbots (*JayBot*) to provide detailed responses to questions about UK universities (Odede, 2024). Similarly, TA chatbots provide personalized support to computer science students, offering guidance and assistance without providing direct answers to their assignments (Liu & M'hiri, 2024).

4. METHODOLOGY

Our research will develop an RAG system using LangChain to improve the integration of LLMs with external data sources. We will also employ the Ragas framework, a novel tool for evaluating RAG systems, to assess our RAG pipeline. Furthermore, we will explore a unique application of RAG within an Internship Placement System, a context not previously explored in the literature. This approach aims to explore the capabilities of RAG implementation in academia to support students' success, contributing to academic and practical advancements in the field.

User Requirement

The approach begins by identifying the user requirements for the internship placement system. These requirements, including the need for seamless access to internship information, personalized assistance, and efficient navigation, were derived from a review of relevant journals and literature, which focus on enhancing the process of obtaining internships for students (Hang et al., 2024; Menezes et al., 2022). Additionally, insights were gained from understanding the challenges faced by our classmates in securing internships. Understanding these requirements is critical for designing a system that caters to the needs of students seeking internships.

Design

The system architecture and design are developed based on user requirements to integrate LLM and RAG into the internship placement system. Both systems involve the development of a chatbot interface capable of interacting with users, understanding their inputs, and providing appropriate responses using the ChatGPT API. FastAPI is hosted on an EC2 instance on AWS with two endpoints, "GPT" and "intern." The "gpt" endpoint receives the user query, adds the prompt, sends it to the OpenAI server, and returns the response (Figure 1). The "intern" endpoint uses the LangChain framework, which adds an information retrieval component. Initially, it gathers data from a new data source using user input. Subsequently, the user's query, prompt, and relevant information are sent to the OpenAI server, and the response is returned (Figure 2).

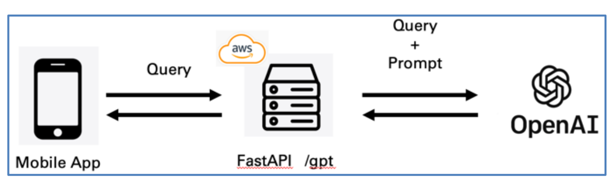


Figure 1: LLM Software Architecture

ISSN: 2473-4901
v10 n6132

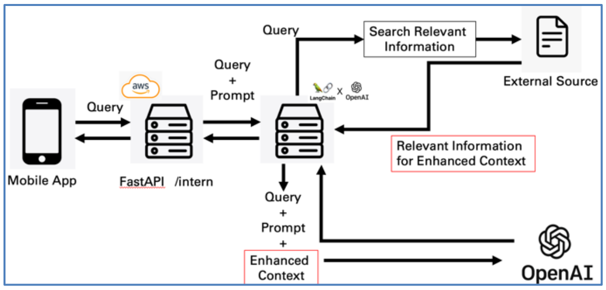


Figure 2: RAG Software Architecture

In addition, for containerization purposes, GitHub Codespaces was used, and for Continuous Integration and Continuous Development (CI/CD) environment, GitHub Actions were employed (Figure 3) (GitHub. n.d.) This automation tool facilitates various stages of the development workflow by enabling automated testing, building, and deployment processes. During the development of the chatbot demo, this automation for DevOps significantly enhanced the overall efficiency and reliability of the development cycle by minimizing manual intervention and enabling faster iterations.

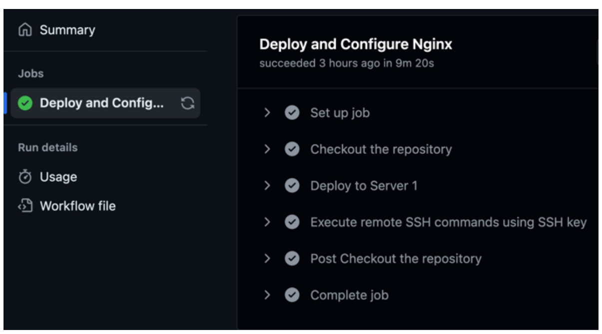


Figure 3: GitHub Actions for CI/CD Pipeline

Implementation

The internship placement system uses a variety of technologies to achieve its goals. The chatbot mobile application is developed using React Native (version 0.74.1), which provides a cross-platform solution for deployment on both iOS and Android devices (React Native RSS, n.d.) (Figure 4). Python (version 3.12.3) and FastAPI (version 0.105.0) serve as the backend framework, providing a robust and efficient environment for handling server-side logic, API creation, and database interactions. FastAPI supports non-blocking I/O operations, leveraging the asynchronous capabilities of Python's asyncio library to handle many requests concurrently and improve performance (FastAPI, n.d.). FastAPI is hosted on an Ubuntu EC2 instance on AWS (Figure 5). The ChatGPT API (GPT-4) is the core component for NLP, allowing the chatbot to effectively understand and respond to user inquiries (OpenAI, n.d.). Furthermore, LangChain (version 0.1.10) supplements the system's functionality by retrieving internship-related information from files and databases, improving the overall user experience (LangChain, n.d.).

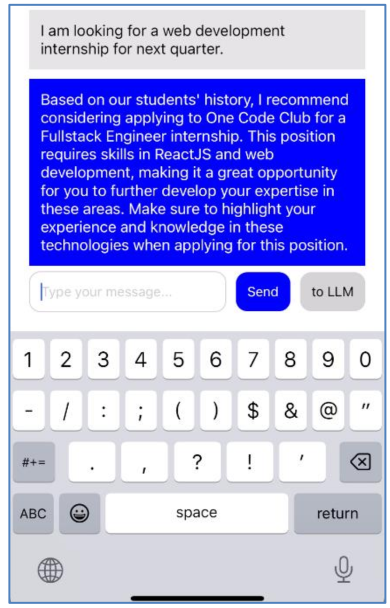


Figure 4: User Interface

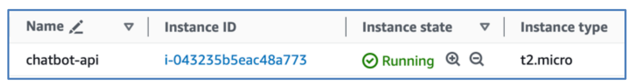


Figure 5: EC2 Instance on AWS

5. Data Collection

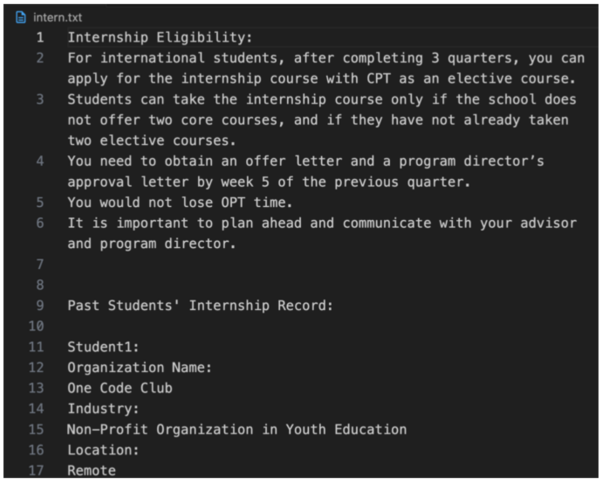
Input Data

User Queries: Simulated user queries about CS internship opportunities are used. These queries are similar to the questions that CS students at a university might ask when seeking information on internship opportunities in their field. For example, "How do I apply for the internship course?" "I have completed a term at City University of Seattle. Am I eligible to apply for the internship course?"

Internship Descriptions: Sample internship descriptions or details are collected from various industries and fields, including information from past students' internship experiences, which are shown in Figure ~~5~~6. These descriptions, tailored specifically for CS students, serve as input data for the LangChain component of the chatbot. The LangChain component plays a crucial role in processing these descriptions, enabling the chatbot to provide relevant responses to help

 ISSN: 2473-4901
v10 n6132

students get internship opportunities. Importantly, the file accessible to the RAG system contains only general information about the university's CS students, ensuring that specific student data remains protected and maintaining student privacy.



```
intern.txt
Internship Eligibility:
For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course.
Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses.
You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter.
You would not lose OPT time.
It is important to plan ahead and communicate with your advisor and program director.


Past Students' Internship Record:

Student1:
Organization Name:
One Code Club
Industry:
Non-Profit Organization in Youth Education
Location:
Remote
```

Figure ~~5~~6: A Sample Internship Description

Output Data

Chatbot Responses: The responses that the LLM system and the RAG system chatbot produced to the same questions from users. Table 1 presents the responses of the LLM and RAG systems to the same set of questions.

6. DATA ANALYSIS

We compare the specificity and relevance of the responses generated by each system. The LLM system's responses in Table 1 were not customized to the unique context of the students at the City University of Seattle; instead, they were generic. The general information in these responses can be applied to a wide range of institutions and contexts. On the other hand, responses from the RAG system in Table 1, which uses LangChain for information retrieval, were noticeably more precise. The RAG responses were customized to meet the unique requirements and circumstances of the City University of Seattle students.

To quantitatively evaluate the performance of these RAG pipelines, we employed the RAG Assessment (Ragas) framework, which helps assess the effectiveness of our RAG systems. While tools and frameworks exist for building these pipelines, quantifying their performance can be challenging. In order to ensure that the chatbot provides users with accurate and pertinent information, Ragas offers metrics and methods for assessing and quantifying the efficacy of our RAG pipelines

Response	Q. How do I apply for the internship course?
LLM	"To apply for the internship course, you typically need to follow these steps: 1. Visit the official website of the organization…"
RAG	"To apply for the internship course at City University of Seattle you need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. It is important to plan ahead and communicate with your advisor and program director."
	Q. I have completed 1 term at City University of Seattle. Am I eligible to apply for the internship course?
LLM	"To determine your eligibility to take the internship course at City University of Seattle, it's best to refer to the specific requirements set by the university or your academic advisor…"
RAG	"For international students at City University of Seattle, you can apply for the internship course with CPT as an elective course after completing 3 quarters. It is important to plan ahead and communicate with your advisor and program director."
	Q. What's the name of the most recently completed internship by a student?
LLM	"I'm sorry, I am not able to provide real-time information about specific individuals or their activities. If you have any other questions or need assistance with general information, feel free to ask!"
RAG	"The most recently completed internship by a student was at One Code Club."

Table 1: Chatbot Response Comparison

The evaluation process involves running predefined queries against the system to compare the generated responses with ground truth answers, calculating key metrics such as context precision, context recall, faithfulness, and answer relevancy using the framework (Ragas, n.d.). Table 2 in the Appendix shows the metric values for three questions asked to the chatbot demo.

ISSN: 2473-4901
v10 n6132

Context Precision metric assesses how relevant the retrieved context is to the question, implying the effectiveness of the retrieval process . High context precision indicates that the context given to the model is highly pertinent to the question, guiding the generation of appropriate answers (Ragas, n.d.). In the application demo, the average context precision stood at 0.999.

The Context Recall metric assesses the retriever's ability to compile all the data required to respond to the query. It ensures the model can access all relevant information needed to respond thoroughly. A high context recall value indicates that the retrieval system effectively collects sufficient context for the given question (Ragas, n.d.). In the application demo, the average context recall stood at 1.0.

Faithfulness metric evaluates the degree of factual consistency of the answer with the context related to the question. It ensures that responses from the chatbot are precise and appropriately matched to the provided context. High faithfulness denotes the answers' credibility and lack of errors or fabrications (Ragas, n.d.). In the application demo, the average faithfulness stood at 0.8889.

Answer Relevancy metric assesses the relevance of the response to the query. It ensures that the answers are accurate and directly relevant to the questions posed by the users. A high answer relevancy indicates that the responses effectively address the users' questions. In the application demo, the average answer relevancy stood at 0.9266.

7. FINDINGS

The data analysis revealed the chatbot's performance across all evaluated metrics, including Context Precision, Context Recall, Faithfulness, and Answer Relevancy. The RAG system, leveraging LangChain for retrieving external data, consistently provided more context-specific answers tailored to the specific needs of City University of Seattle users. In comparison, the LLM system generated more generic responses, which could be applied to a larger variety of users.

Throughout the assessment, these metrics, each representing a different aspect of the chatbot's functionality, repeatedly produced high numbers close to 1.0. High Context Precision indicates the model generated appropriate and accurate responses, as the retrieved context was highly relevant to the users' questions. Similarly, a high Context Recall value suggests that the retrieval system successfully gathered all data needed to offer comprehensive answers, ensuring comprehensive responses to user queries.

Furthermore, the chatbot's responses were consistent with the context given, as indicated by the high Faithfulness, which promotes confidence in the accuracy of the information provided. In addition, the high Answer Relevancy indicates that the chatbot regularly provided directly relevant answers to the users' queries. This performance can be associated with the chatbot's use of LangChain, which empowered the ability to understand and respond to user-specific queries precisely and accurately.

8. CONCLUSIONS

Overall, the availability of NLP-powered services, such as OpenAI API and LangChain, has provided numerous opportunities for developers across multiple domains. The integration of LLMs and RAG into the internship placement system exemplifies this. It demonstrates the efficacy of using LLM and RAG to provide personalized chatbot services to students.

By evaluating key metrics such as Context Precision, Context Recall, Faithfulness, and Answer Relevancy, we have observed high levels of effectiveness in the RAG approach in providing accurate and relevant responses to user queries. Utilizing the power of LLM and RAG, the application can provide tailored responses to student inquiries, improving the user experience and facilitating seamless interactions within the system.

Our problem statement aimed to determine whether RAG can provide a better user experience compared to LLMs. These findings suggest that the adoption of cutting-edge technologies such as LLM and RAG has the potential to revolutionize NLP applications like chatbots. This would stimulate innovation and enhance user experiences for both individuals and enterprises.

9. VALIDATIONS

The validation of our research results and the chatbot demo performance experienced certain constraints due to the ten-week capstone course's limited duration. The time constraint significantly restricted our ability to develop and test the optimal RAG pipeline. Testing was conducted by simulating user inputs and evaluating the RAG pipeline using the RAGAS

ISSN: 2473-4901
v10 n6132

framework. However, accurate user testing would have provided valuable insights into user patterns and behavior, allowing for enhancements to the application and additional findings to improve the RAG system. Furthermore, the size and scope of our RAG input data were constrained. With a more extensive and diverse dataset, we could have performed more extensive and rigorous testing, allowing for a more comprehensive evaluation of the RAG pipeline.

10. FUTURE WORK

Future work should expand the RAG pipeline's training and testing dataset. By integrating a more extensive array of internship descriptions and user inquiries, it is possible to effectively verify the chatbot's capacity to provide accurate and relevant data. Further optimization and enhanced performance of the RAG pipeline will likely result from longer-term iterative testing and refinement. Actual user tests must be included to obtain feedback and insights from real users. It will allow a more realistic evaluation of the chatbot's performance and user experience.

11. REFERENCES

Abdullah, F. S., Abu Seman, A., Ibrahim, N., Majid, N. A., Abdul Wahab, N. M., Mustafa, M. S., & Mohd Sani, N. F. (2017). Web-based application of the internship management system. Journal of Computing Research and Innovation (JCRINN), 2(3), 46-51.

Chaurasia, S. (2023). Student Internship Placement Management System using python. April-May 2023, (33), 30–49. https://doi.org/10.55529/ijrise.33.30.49

FASTAPI. FastAPI. (n.d.). https://fastapi.tiangolo.com/

Gautam, B., & Purwar, A. (2024). Evaluating the efficacy of open-source LLMs in enterprise-specific RAG systems: A comparative study of performance and scalability. arXiv. https://arxiv.org/abs/2406.11424

GitHub. (n.d.). Secure development made simple. https://github.com/features/codespaces

Hang, K., Schrock, T., Ostrander, T. J., Berg, R., Menezes, T., & Wang, K. (2024). Industry mentoring and internship experiences at a Community College Baccalaureate Program in Software Development. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1. https://doi.org/10.1145/3626252.3630878

Langchain. LangChain. (n.d.). https://www.langchain.com/

Liu, M., & M'Hiri, F. (2024). Beyond traditional teaching: Large language models as simulated teaching assistants in Computer Science. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. https://doi.org/10.1145/3626252.3630789

Martineau, K. (2024, May 1). What is retrieval-augmented generation (rag)?. IBM Research. https://research.ibm.com/blog/retrieval-augmented-generation-RAG

Menezes, T., Parra, A., & Jiang, M. (2022). Open-source internships with industry mentors. *Proceedings of the 27th ACM Conference on*

Odede, J., & Frommholz, I. (2024a). JayBot -- Aiding University students and admission with an LLM-based chatbot. *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. https://doi.org/10.1145/3627508.3638293

OpenAI. (n.d.). https://openai.com/

Ragas. (n.d.). Introduction. https://docs.ragas.io/en/stable/

React native. React Native RSS. (n.d.). https://reactnative.dev/

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of LLMS in practice: A survey on CHATGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/3649506

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. Retrieved May 20, 2024, from https://arxiv.org/pdf/2402.19473

ISSN: 2473-4901
v10 n6132

Appendix

Question 1	I have completed 1 term at City University of Seattle. Am I eligible to apply for the internship course?			
Answer	"Based on the provided context, for international students at City University of Seattle, you can apply for the internship course with CPT as an elective course after completing 3 quarters. It is important to plan ahead and communicate with your advisor and program director."			
Context	"Internship Eligibility: For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director."			
Ground Truth	You are eligible to apply for the internship course after completing 3 quarters at City University of Seattle			
	context_precision	context_recall	faithfulness	answer_relevancy
	1.0	1.0	0.6667	0.9051
Question 2	How do I apply for the internship course?			
Answer	To apply for the internship course at City University of Seattle, you need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. It is important to plan ahead and communicate with your advisor and program director.			
Context	"Internship Eligibility: For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director."			
Ground Truth	You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter.			
	context_precision	context_recall	faithfulness	answer_relevancy
	0.9999	1.0	1.0	0.8965
Question 3	What is the name of the most recently completed internship by a student?			
Answer	The most recently completed internship by a student was at One Code Club in the 2023 Spring Quarter, where the student worked as a Full Stack Engineer on frontend development of a website using Next.js.			
Context	"Internship Eligibility:			

 ISSN: 2473-4901
v10 n6132

	For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director’s approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director.”			
Ground Truth	One Code Club			
	context_precision	context_recall	faithfulness	answer_relevancy
	0.9999	1.0	1.0	0.9783

Table 2: RAG Assessment Values