



**Department of Electrical and Computer Engineering  
North South University**

## **Directed Research**

# **Leveraging LLMs on Root Level Education System**

<b>Abraham Kaikobad</b>	<b>ID # 2013983042</b>
<b>Md Nazmus Sadat Samin</b>	<b>ID # 2012235042</b>
<b>Tanjila Ahmed Medha</b>	<b>ID # 2021004042</b>
<b>Rian Ahmed</b>	<b>ID # 2013650042</b>

### **Faculty Advisor:**

**Dr. Shahnewaz Siddique**

**Associate Professor**

**Department of Electrical and Computer Engineering**

**Spring, 2024**

## APPROVAL

Abraham Kaikobad (ID # 2013983042), Md Nazmus Sadat Samin (ID # 2012235042), Tanjila Ahmed Medha (ID # 2021004042) and Rian Ahmed (ID # 2013650042) from Electrical and Computer Engineering Department of North South University, have worked on the Directed Research Project titled “**Leveraging LLMs on Root Level Education System**” under the supervision of Dr. Shahnewaz Siddique partial fulfilment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

### Supervisor’s Signature

.....

**Dr. Shahnewaz Siddique**

**Associate Professor**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

### Chairman’s Signature

.....

**Dr. Rajesh Palit**

**Professor**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

## DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

### Students' names & Signatures

1. Abraham Kaikobad

.....

2. Md Nazmus Sadat Samin

.....

3. Tanjila Ahmed Medha

.....

4. Rian Ahmed

.....

## ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Shahnewaz Siddique, Associate Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice about the experiments, research, and theoretical studies carried out during the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

## ABSTRACT

# Leveraging LLMs on Root Level Education System

Education is key to personal growth, career success and social development. It empowers individuals with knowledge, critical thinking and problem-solving skills, ensuring them to contribute to their communities and worldwide. From Bangladesh's perspective, there are some challenges in the education system. The traditional education system has some drawbacks like lack of individual concentration, inflexible learning, limited resources on books, assessment of teaching effectiveness and high costs. With the rise of Artificial Intelligence in our daily lives, it will be a great prospect for our education system. Integrating Large Language Models(LLMs) into our education system can transform teaching and learning methods by providing personalized feedback, multilingual adaption, content accessibility, education content creation and improved assessment facility. LLms can create captivating learning environments. Therefore, realizing the prospect of LLMs in the education system, we leveraged state-of-the-art large language models and frameworks, LLamA2 and LLamAIndex to create a knowledgable QA assistant to easily navigate a vast number of PDF documents. We developed a Retrieval-Augmented Generation (RAG) system using the LLamA2 model from Hugging Face. With LLamA2's proficient language understanding and LLamAIndex's effective information retrieval skills combined, we intend to build a system that can precisely respond to queries.

**Keywords :** Large Language Models, RAG, LlamA2, Llamaindex

# Contents

<b>APPROVAL</b>	<b>i</b>
<b>DECLARATION</b>	<b>ii</b>
<b>ACKNOWLEDGEMENTS</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>LIST OF FIGURES</b>	<b>vi</b>
<b>LIST OF TABLES</b>	<b>vii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Purpose and Goal of the Project . . . . .	2
1.3 Organization of the Report . . . . .	2
<b>2 Research Literature Review</b>	<b>3</b>
2.1 Existing Research and Limitations . . . . .	3
<b>3 METHODOLOGY</b>	<b>5</b>
3.1 System Design . . . . .	5
3.2 Software Components . . . . .	6
3.3 Software Implementation . . . . .	7
<b>4 Investigation/Experiment, Result, Analysis and Discussion</b>	<b>8</b>
4.1 Investigation/Experiment . . . . .	8
4.2 Results and Analysis . . . . .	9
<b>5 CONCLUSION</b>	<b>11</b>
5.1 Summary . . . . .	11
5.2 Limitations . . . . .	11
5.3 Future Improvement . . . . .	12
<b>BIBLIOGRAPHY</b>	<b>14</b>

## List of Figures

3.1	A high-level overview of the Retrieval-Augmented Generation (RAG) methodology presented in this study. The process involves splitting generated responses into semantic chunks, storing them in a vector-based knowledge base, and then performing intelligent retrieval and generation to provide high-quality, contextually relevant answers to user questions. . . . .	5
-----	---	---

## List of Tables

4.1	Evaluation outputs from the PDFs . . . . .	10
-----	--	----



# Chapter 1

## INTRODUCTION

### 1.1 Background and Motivation

Bangladesh's education system has several drawbacks, including low Quality of Education and inequitable Access. Poor teacher quality leads to significant learning losses over eleven years of schooling. The system is criticized for its outdated curriculum and lack of relevance to practical life and teaching methods that emphasize rote learning over critical thinking. In contrast to first-world countries, Bangladesh's education system lacks the resources, infrastructure, and modern teaching methods that are common in developed nations.[1] If we want to compete with the developed countries around the world we need to integrate modern technology and its application in our root-level education system. The integration of Large Language Models (LLMs) like Llama2 and LlamaIndex into the education system holds immense potential to revolutionize the way students access and engage with information. The goal of this PDF-based QA Assistant project is to provide an intelligent and user-friendly system that enables students to easily navigate a vast repository of educational materials with ease. This project's motivation is the increasing need for effective, individualized education in the 21st century. Conventional teaching approaches frequently find it difficult to keep up with students' learning demands and the amount of knowledge that is always growing. This approach attempts to close the knowledge gap between students and the vast amount of information included in PDF documents by utilizing the power of Retrieval-Augmented Generation (RAG) techniques. This project is based on Meta's Llama2 model[2], which provides the natural language processing capabilities crucial to enable smooth communication between students and the system. A data framework for LLM applications called LlamaIndex makes it easier to retrieve, organize, and input appropriate information from PDF documents, enabling students to obtain the answers they need. The design and execution of this project has been advised and inspired by multiple notable research papers, including "Rethinking search: making domain experts out of dilettantes" by Metzler et al[3] and "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" by Lewis et al[4], which explores the integration of retrieval mechanisms with language models. This PDF-based QA Assistant aims to provide students with a personalized and engaging learning experience, establishing an in-depth knowledge of course materials and increasing

academic success by integrating the benefits of Llama2 and LlamaIndex.

## **1.2 Purpose and Goal of the Project**

The purpose of this project is to revolutionize the educational landscape in Bangladesh by integrating Large Language Models (LLMs) into the root-level education system. The primary goal is to enhance student understanding and assessment by leveraging the advanced capabilities of LLMs. By developing a PDF-based QA Assistant using Llama2 and LlamaIndex, we aim to create a comprehensive and accessible knowledge repository that students can utilize to improve their understanding of course materials. This system will enable students to get vast amounts of information efficiently, fostering a deeper and more nuanced understanding of subjects. Additionally, the project aims to empower teachers by providing them with a robust tool to create high-quality educational content. This will not only improve the quality of teaching but also enhance the overall learning experience for students. By integrating LLMs into the educational ecosystem, we envision a future where students are better equipped to tackle complex problems, think critically, and develop essential skills for the 21st century.

## **1.3 Organization of the Report**

The structure of this report is as follows: the study that has already been done and its limits on the project; an explanation of how PDF-based QA assistant. What is the Large Language model, and how does it work in speech-related tasks? A thorough description of each component in the suggested model and the stages and procedures used in the experiments. What is the real-world impact of this project? The project's financial planning and planning methodology After that, the topic of complicated engineering tasks and challenges was covered. Finally, the conclusion, its shortcomings, and the work that has to be done

## Chapter 2

### Research Literature Review

#### 2.1 Existing Research and Limitations

The integration of Large Language Models (LLMs) into the education system has been a topic of growing interest, with researchers exploring the potential of these advanced language models to revolutionize the way students learn and teachers deliver content. One notable study, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", [4] investigates the use of retrieval mechanisms in conjunction with language models to enhance the performance of knowledge-intensive tasks. The authors demonstrate how the combination of retrieval and generation can lead to significant improvements in areas such as question-answering and summarization. This research lays the foundation for the development of our PDF-based QA Assistant, which aims to leverage the power of Retrieval-Augmented Generation (RAG) to provide students with personalized and efficient access to educational resources. The paper "Evaluating the Retrieval Component in LLM-Based Question Answering Systems" proposes a framework for assessing the retrieval component in Retrieval-Augmented Generation (RAG)-based chatbots. [5] They developed a straightforward baseline for assessing the performance of retrievers in LLM-based question-answering systems. Their framework is more aligned with the actual performance of the QA system compared to traditional retrieval metrics like precision and recall. "ChatGPT and large language models in academia: opportunities and challenges" by Meyer et al [6] explores the integration of ChatGPT and large language models (LLMs) in academic settings. Then in 2023, Karl de Fine Licht offers a framework for incorporating LLMs into higher education in his research named, "Integrating Large Language Models into Higher Education: Guidelines for Effective Implementation". [7]. The study discusses practical applications of LLMs in education, such as generating text and assisting with research tasks, to enhance academic work efficiency. He offers valuable insights for integrating Learning and Learning Management (LLMs) into higher education to enhance learning outcomes and promote innovation in educational practices. The paper "Large Language Models in Education: Vision and Opportunities" [8] explores the potential of large language models (LLMs) to transform the education sector by addressing the limitations of traditional education systems. The

paper presents a compelling vision for the integration of LLMs in education, highlighting the immense potential to improve educational outcomes and pave the way for a more engaging and effective learning environment.

The papers reviewed in the literature highlight several key limitations that must be addressed to effectively integrate Large Language Models (LLMs) into the education system. One of the primary concerns is the issue of data privacy and security, as the integration of LLMs into educational settings raises questions about the protection of sensitive student data. Ethical considerations are also crucial, as the potential for bias and the need for transparency in model development and deployment must be carefully navigated to maintain the legitimacy and fairness of the system. Another significant limitation is the infrastructure and resource constraints that often accompany the implementation of LLMs. Ensuring that schools and educational institutions have the necessary computational resources and infrastructure to effectively deploy and utilize these advanced language models is a critical challenge. Multilingual support and accessibility are also important considerations, as the education system must cater to students from diverse linguistic backgrounds. Developing LLMs that can effectively translate and adapt content to meet the needs of students from different language backgrounds is essential for providing equitable access to educational resources. Finally, the integration of LLMs into existing educational systems presents its own set of challenges, as the technology must be seamlessly incorporated into different workflows and infrastructures. Addressing these integration challenges is crucial for ensuring that the implementation of LLMs in the education system is smooth and effective.

By addressing these limitations, the integration of LLMs into the education system can be optimized to provide a more personalized, efficient, and effective learning experience for students, ultimately contributing to the betterment of the educational landscape.

## Chapter 3

### METHODOLOGY

#### 3.1 System Design

The development of artificial intelligence (AI) and machine learning has revolutionized how humans interact with information, making it more accessible, comprehensible, and usable. This study demonstrates the construction of an intelligent Q&A assistant leveraging cutting-edge language models and indexing frameworks. By combining large language models (LLMs) like LLaMA2 and indexing tools like LLaMAIndex, we present a Retrieval-Augmented Generation (RAG) system capable of efficiently browsing and answering queries across a vast corpus of PDF documents.

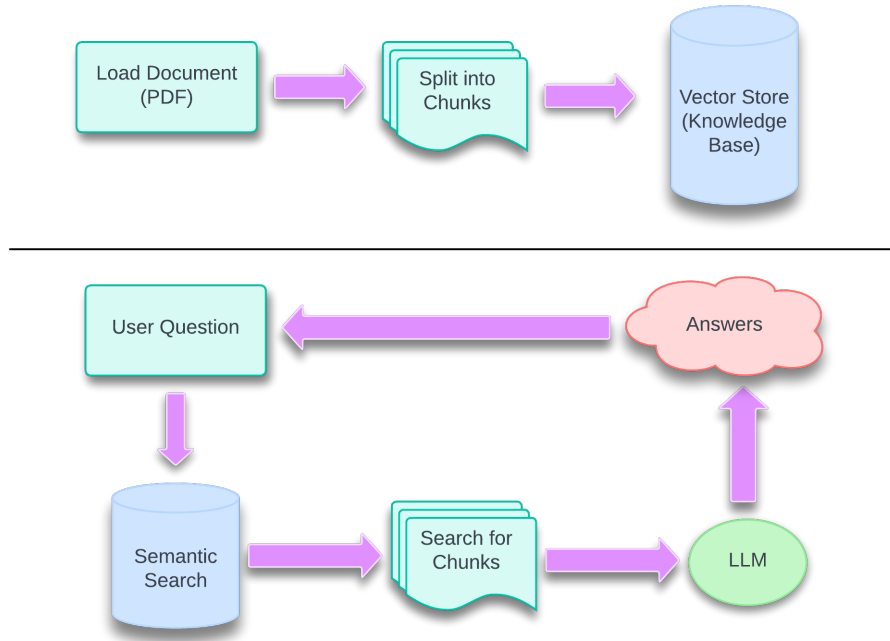


Figure 3.1: A high-level overview of the Retrieval-Augmented Generation (RAG) methodology presented in this study. The process involves splitting generated responses into semantic chunks, storing them in a vector-based knowledge base, and then performing intelligent retrieval and generation to provide high-quality, contextually relevant answers to user questions.

Figure 3.1 describes that developers, data scientists, and tech enthusiasts build natural language processing systems that rival industry leaders. The critical aspects of this approach

include splitting generated responses into semantic chunks, storing them in a vector-based knowledge base, and performing intelligent retrieval and generation to provide high-quality, contextually relevant answers to user questions[1]. This tutorial teaches readers how to construct a powerful Q&A assistant that leverages the latest advancements in AI and language modeling.

## 3.2 Software Components

This study challenges theoretical ideas and their helpful use in real-world situations as part of our attempt to demystify the development of an AI-driven Q&A assistant. We aim to build a system that provides accurate responses and broadens our knowledge of the opportunities and difficulties in natural language processing (NLP) by combining LLama2's sophisticated language comprehension with LLamaIndex's effective information retrieval capabilities.

**LLama2 Model.** LLaMA2, a new version of the LLaMA language model, has outperformed its predecessor, LLaMA version 1, released in July 2023. It comes in three sizes: 7B, 13B, and 70B parameter models. Upon its release, LLaMA 2 achieved the highest score on Hugging Face, with the top-performing model originating from LLaMA 2. Llama 2 was trained on 2 Trillion Pretraining Tokens and has outperformed state-of-the-art open-source models like Falcon and MPT in various benchmarks. It also underwent fine-tuning for chat-related use cases, training with over 1 million human annotations. LLama2 is a beacon of innovation in natural language processing, pushing the boundaries of what's possible with language models. Its architecture is designed for efficiency and effectiveness, allowing for an unprecedented understanding and generation of human-like text. Unlike its predecessors, LLama2 offers a more nuanced approach to processing language, making it particularly adept at tasks requiring deep comprehension.

**LLamaIndex.** Indexing is the foundation of every effective information retrieval system. LLamaIndex, a framework for document indexing and querying, distinguishes itself by offering a streamlined approach to managing extensive collections of documents. It is not enough to just store information; it must also be accessible and retrievable instantly. The relevance of LLamaIndex cannot be emphasized, as it allows for real-time query processing across large databases, guaranteeing that our Q&A assistant can deliver rapid and accurate replies based on a rich knowledge base.

**Tokenization and Embeddings.** The first step in understanding language models involves breaking down text into manageable pieces, a process known as tokenization. This foundational task is crucial for preparing data for further processing. Following tokenization, the concept of embeddings comes into play, translating words and sentences into numerical vectors.

**Model Quantization.** Model quantization presents a strategy to enhance the performance and efficiency of our Q&A assistant. By reducing the precision of the model's numerical computations, we can significantly decrease its size and speed up inference times. While introducing a trade-off between precision and efficiency, this process is precious in resource-constrained environments such as mobile devices or web applications. Through careful application, quantization allows us to maintain high levels of accuracy while benefiting from reduced latency and storage requirements.

**ServiceContext and Query Engine.** The ServiceContext within LLamAIndex is a central hub for managing resources and configurations, ensuring that our system operates smoothly and efficiently. The glue holds our application together, enabling seamless integration between the LLamA2 model, the embedding process, and the indexed documents. On the other hand, the query engine is the workhorse that processes user queries, leveraging the indexed data to fetch relevant information swiftly. This dual setup ensures that our Q&A assistant can easily handle complex queries, providing quick and accurate answers to users.

### 3.3 Software Implementation

The LLamA2 model is initialized with specific parameters for a Q&A system, demonstrating its versatility and adaptability to various contexts and applications. The embedding model is crucial for capturing the semantic essence of documents, and Sentence Transformers are used to gauge textual content similarity and relevance. The ServiceContext is instantiated with default settings, ensuring all system components are harmonized and ready for indexing and querying operations. The setup transitions data preparation to actionable insights, enabling the Q&A assistant to respond to queries based on indexed content. The system is tested by querying for summaries and insights from the document collection, demonstrating the practical utility of the Q&A assistant and the seamless integration of LLamA2, LLamAIndex, and underlying NLP technologies.

## Chapter 4

### Investigation/Experiment, Result, Analysis and Discussion

#### 4.1 Investigation/Experiment

In this project, we aimed to develop a PDF-based Q&A Assistant that leverages the power of the Llama2 language model to provide students with efficient and personalized access to educational resources. To achieve this, we utilized a dataset comprising class 6 and 7 English, Math, and Science textbooks, which served as the foundation for our knowledge base.

The critical components of our project setup are as follows: **Dataset Preparation.** We carefully selected a dataset of PDF files containing the class 6 and 7 textbooks for English, Math, and Science subjects. These textbooks were chosen as they represent the core curriculum at the root level of the education system, ensuring that the Q&A Assistant can address the fundamental learning needs of students.

**Llama2 Model Integration:** We selected to use Hugging Face’s Llama2 language model as the foundation of our Q&A Assistant. Llama2 is a big language model that has been pre-trained on massive data, enabling it to decode and generate natural language with excellent accuracy.

**LlamaIndex Integration.** We implemented the LlamaIndex framework to enable effective information retrieval from the PDF dataset. A data architecture called LlamaIndex was explicitly created for LLM applications. It offers a smooth method for ingesting, organizing, and querying massive volumes of textual data.

**Retrieval-Augmented Generation (RAG).** To take advantage of Llama2’s capabilities, we created a Retrieval-Augmented Generation (RAG) system. This method combines the benefits of information retrieval and building languages, allowing the Q&A Assistant to immediately retrieve relevant information from the PDF collection and respond constructively and effectively to student queries.

The workflow of our project starts with data ingestion into the Llama-index framework. The class 6 and 7 textbook PDFs are ingested into the LlamaIndex framework, which creates a structured representation of the content. Then, a student submits a question, the Q&A Assistant uses the LlamaIndex to retrieve the most relevant passages from the textbook data. The Llama2



model then processes the recovered passages to provide a logical and informative response to the student's inquiry. The generated response is displayed to the output cell, providing users with the necessary information to enhance their understanding of the subject matter.

**Configurations.** The use of Google Colab Pro has been instrumental in the development of our Llama2-based PDF-based Q&A Assistant project. The Pro version of this cloud-based platform offers access to high-performance computing resources, including the powerful NVIDIA A100 GPU, which has been crucial for our project. The A100 GPU provides exceptional processing power, allowing us to efficiently train and run the large Llama2 language model on the extensive PDF dataset. The 25GB of RAM offered by the Pro version further supports the handling of the Llama2 model and the processing of the textbook data. Additionally, the ample 100GB of disk space and extended runtime limits enable us to store the necessary files and run our Q&A Assistant without interruption. By leveraging the resources available in Google Colab Pro, we have been able to develop a highly efficient and responsive system that delivers accurate and informative responses to student queries.

## 4.2 Results and Analysis

In this study, The LLamA2 model is initialized with specific parameters for a QA system, demonstrating its versatility and adaptability to various contexts and applications. The embedding model is crucial for capturing the semantic essence of documents, and Sentence Transformers are used to gauge textual content similarity and relevance. The ServiceContext is instantiated with default settings, ensuring all system components are harmonized and ready for indexing and querying operations. The setup transitions data preparation to actionable insights, enabling the QA assistant to respond to queries based on indexed content. The system is tested by querying for summaries and insights from the document collection, demonstrating the practical utility of the Q&A assistant and the seamless integration of LLamA2, LLamAIndex, and underlying NLP technologies.

Table 4.1: Evaluation outputs from the PDFs

Giving Prompt	Corresponding Answers
In what ways is an ant's life peaceful?	<p>Based on the given text, an ant's life is peaceful in several ways:</p> <ol style="list-style-type: none"> <li>1. Ants do not fight with each other or with other members of their group. They work together intelligently and bravely and never harm each other.</li> <li>2. Ants have a well-organized society with separate roles for soldiers, workers, and cleaners. This means there is no competition or conflict among them.</li> <li>3. Ants spend most of their time searching for food but do not fight with other ants or creatures for resources. They have a system of storing food in designated areas, which ensures that there is always enough food available for everyone.</li> </ol> <p>Therefore, an ant's life is peaceful because of the lack of conflict, competition, and struggle for survival. They work together harmoniously and have a well-organized society that allows them to thrive in their environment.</p>
How many thousands are in 84806?	<p>To find out how many thousands are in 84,806, you divide 84,806 by 1,000:</p> $84806 \div 1000 = 84.806$ <p>So, there are 84 thousand in 84,806.</p>
Can you guess why algae are green in color?	<p>Algae are green in color primarily due to chlorophyll pigments within their cells. Chlorophyll is a vital pigment that absorbs light energy, particularly in the blue and red wavelengths of the electromagnetic spectrum, while reflecting green light. This phenomenon gives algae their characteristic green coloration.</p> <p>While chlorophyll is the primary pigment responsible for algae's green color, different types of algae may also contain accessory pigments such as carotenoids and phycobilins, which can modify their shade of green or even produce colors like red or brown in certain species. These variations in pigment composition help algae adapt to different light conditions and ecological niches, contributing to their diverse range of colors observed in nature.</p>

## Chapter 5

### CONCLUSION

#### 5.1 Summary

This study underscores the importance of aligning large language models (LLMs) with human values to ensure trustworthiness, reliability, ethical decision-making, and user safety. By introducing a solution for managing information overload by creating customized chatbots powered by large language models (LLMs). It highlights the capability of summarizing PDF documents and providing answers to user questions from uploaded PDFs, enhancing productivity and efficiency in information retrieval. Key components include OpenAI for language understanding, LangChain for context-aware reasoning, and Streamlit for building interactive web applications. The purpose of this report is to guide new users through setting up libraries, extracting text from PDFs, segmenting text, generating embeddings, and connecting with LLamA2 for Q/A capabilities.

#### 5.2 Limitations

While the PDF-based Q&A Assistant leveraging LLama2 and LlamaIndex presents a promising approach to enhancing the educational experience, there are several limitations that must be addressed to ensure its long-term effectiveness. One key concern is the potential for data bias and inaccuracies within the PDF dataset used to train the system, which can significantly impact the quality and relevance of the responses provided to students. Additionally, the current system may be limited in its ability to support students from diverse linguistic backgrounds, highlighting the need for robust multilingual capabilities. Personalization and adaptive learning features could also be integrated to tailor the content and delivery to individual student needs, fostering a more engaging and effective learning experience. Finally, seamless integration with existing educational infrastructure, such as learning management systems and student information databases, would enable a more holistic and streamlined approach to leveraging the Q&A Assistant within the broader educational ecosystem. By addressing these limitations through techniques like active learning, multi-modal data integration, and collaborative development

with educators, the project can work towards transforming the root-level education system.

### **5.3 Future Improvement**

To enhance our PDF-based QA Assistant, key areas for improvement include expanding multi-modal capabilities to handle diverse content formats, implementing personalized learning features to cater to individual needs, improving open-ended question handling and natural conversation abilities, and integrating the system with existing educational platforms. Incorporating these advancements can create a more engaging, adaptive, and seamless learning experience for students. By leveraging techniques like computer vision, adaptive learning, and dialogue management, the QA Assistant can evolve into a powerful tool that significantly improves the learning experience at the root level of the education system. Addressing these areas of improvement will be crucial for the project's future development and impact.

## Bibliography

- [1] A. A. Haider, Problems with our education sector, <https://www.thedailystar.net/problems-with-our-education-sector-23954>, accessed: 2024-7-13 (May 2014).
- [2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models (2023). arXiv:2307.09288.  
URL <https://arxiv.org/abs/2307.09288>
- [3] D. Metzler, Y. Tay, D. Bahri, M. Najork, Rethinking search: making domain experts out of dilettantes, SIGIR Forum 55 (1). doi:10.1145/3476415.3476428.  
URL <https://doi.org/10.1145/3476415.3476428>
- [4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks (2021). arXiv:2005.11401.  
URL <https://arxiv.org/abs/2005.11401>
- [5] A. Alinejad, K. Kumar, A. Vahdat, Evaluating the retrieval component in llm-based question answering systems (2024). arXiv:2406.06458.  
URL <https://arxiv.org/abs/2406.06458>
- [6] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez, J. H. Moore, ChatGPT and large language models in academia: opportunities and challenges, BioData Min. 16 (1) (2023) 20.

- [7] K. de Fine Licht, Integrating large language models into higher education: Guidelines for effective implementation, *Computer Sciences and Mathematics Forum* 8 (1). doi: 10.3390/cmsf2023008065.  
URL <https://www.mdpi.com/2813-0324/8/1/65>
- [8] W. Gan, Z. Qi, J. Wu, J. C.-W. Lin, Large language models in education: Vision and opportunities (2023). arXiv:2311.13160.  
URL <https://arxiv.org/abs/2311.13160>