e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707173 |

Content and Question Generation from Book using LLM, Lang Chain, and FAISS

Karishma M Mujawar, Prof. Vidya S

PG Student, Department of Master of Computer Application, Bangalore Institute of Technology, VVPuram, Banaglore, India

Assistant Professor, Department of Master of Computer Application, Bangalore Institute of Technology, VVPuram, Banaglore, India

ABSTRACT: Creating an application for question generation from PDF documents is challenging so as to achieve this, the content of PDF`s should be evaluated and formulate thoughtful questions. In order to produce new applications, this initiative aimed to automatically construct test question sets for educational evaluations from PDF documents utilizing the powerful natural language processing (NLP) framework Lang Chain [1].

It should be emphasized that the process of formulating questions may eventually become automated. This will likely improve student participation and impression of the subject as well as the change the way instructors behave themselves. In the particular PDFs, the application produced well-considered and acceptable questions across a variety of academic areas and skill levels. These findings show that the quality assessment questions generated by Lang Chain are appropriately evaluated when the additional built-in data is carefully incorporated into the PDFs used in this study [2].

I.INTRODUCTION

The rapid development of artificial intelligence has spawned a plethora of novel concepts and applications in fields such as research, education, and content production. Question generation and content auto-generation from books using Lang Chain, FAISS, and Large Language Models (LLMs) are a few examples.

Multiple choice questions, or MCQs, are now frequently utilized in exams when a large number of students are taking the test, such as in the case of GATE, CAT, NET, and other exams. Since MCQ is very simple to evaluate and is done so using computerized tools, results can be announced in a matter of hours and the evaluation process is 100% clean. Researchers have examined how exams are administered and determined whether or not computer applications can generate exam questions automatically, which would lessen the workload for teachers. NLP is a field of study and application that looks at the practical applications of using computers to read, comprehend, and manipulate voice or text in natural language[1].

NLP has a lot of potential to create computer interfaces that are more user-friendly.

Since individuals will be able to communicate with computers in their own language instead of needing to learn a specialized language of computer commands, natural language processing (NLP) holds enormous potential for creating more user-friendly computer interfaces. In order to build the right tools and strategies to help computer systems understand and manipulate natural languages in order to do desired tasks, NLP researchers seek to acquire information about how humans interpret and utilize language. From the days of punch cards and batch processing to the Google era, NLP research has changed with time.

Up to the present, the creation of summaries of books and generating questions has required a manually intensive step whereas researchers invest a great deal of time in searching for valuable and important knowledge within a large number of written texts. Such manual work is often invoked and creates inconsistency, and the insights are not easily available to other departments. It has become quite evident that there is an urgent need for more of a scalable efficient and most of all consistent solution. An approach was employed to generate questions from the PDF documents in order to implement the research plan [2].

II.LITERATURE REVIEW

The summary includes related to NLP and machine translation in developing an application that can generate questions from PDF documents. This task requires carefully analyzing the content of the PDFs to create meaningful and informative questions. The paper aims to introduce new software that can create test questions automatically for educational purposes from PDF documents using Lang Chain, a powerful NLP Framework [3].

Relevance to current Research

The research method involved several steps to generate questions from PDFs. Initially, NLP techniques were used to extract named entities and their connections from the PDF content. This step involved identifying key information, conceptual relationships, and significant events in the document. The extracted information was then processed using Lang Chain, a probabilistic language model. Leveraging its contextual awareness and understanding of knowledge patterns, Lang Chain generated well-structured and meaningful questions.

The study contributes to the theoretical advancement of methodologies and knowledge representation in NLP. It achieves a precision of 0.74 in Recall-Oriented Understudy for Gisting Evaluation (ROUGE) testing, surpassing other systems that score 0.5. Additionally, it attains an F1 score of 0.88 in BERT Score, outpacing other QA applications with a score of 0.81. In Bilingual Evaluation Understudy (BLEU) testing, RAG achieves a precision of 0.28, significantly more than the 0.09 of others, and it scores 0.33 in Jaccard Similarity, compared to 0.04 by other systems [4].

Relevance to current Research

The study introduces the Retrieval Augmented Generation (RAG) method to enhance Question-Answering (QA) systems, specifically addressing challenges in processing documents within NLP. RAG represents a significant advancement in document question-and-answer applications, overcoming limitations of previous QA systems. By integrating search techniques within a data store and leveraging text generation capabilities of LLMs, RAG offers a more efficient solution compared to traditional manual reading.

The research focuses on evaluating the performance of RAG when using the Generative Pre-trained Transformer 3.5 (GPT-3.5-turbo) from the ChatGPT model. This evaluation is conducted by comparing RAG's effectiveness against other existing applications in processing document data. To test the QA system's capabilities, the study utilizes a newly proposed dataset along with the well-known Stanford Question Answering Dataset (SQuAD) [5].

Limitations of Traditional Methods: Many years ago, information was obtained through reading, including manual reading of documents, papers, and journals among others. Traditional techniques such as using keyword search have been used extensively for extracting information from PDF documents. However, these methods are always so time-consuming, uses a lot of force and most of the time incorrect. Research [1] highlighted the drawbacks of human reading, essentially calling for the automatic reading to be conducted to minimize on them. Similarly studies [2] and [3] highlighted the efficiency of the keyword-based search and compared it to its inability to handle complex searches and extract specific details from the context of the files which are in PDF format.

III.EXSISTING AND PROPOSED SYSTEM

Self-generating content and questions are the use of AI and NLP that help to derive content from books such as summaries, questions etc. Most of the automated content and question generation systems employed in today’s are based on different classes of Machine learning techniques and NLP techniques. Such systems deal with large text volumes, produce valuable information and related queries. Besides, scalability, efficiency, and integration are not very high. The following sections expound the current systems’ pivotal attributes and functionalities prior to implementing the proposed system that consists of LLM, Lang Chain, and FAISS.

It was thus proposed that the key features within the system well be incorporated to enrich the values of the user interfaces and result in better search outcomes from documents. Firstly, this work uses conversational search interface and such an approach is helpful as the users can ask a question. The system focuses on the user end by providing graphic a user-friend land attractive design of the application, providing the needed navigation across documents. These features collectively to become helpful tools for users and provide the means to efficiently facilitate the discovery of documents [6].

Advantages of Proposed System

• Improved retrieval accuracy: Locate the information that is related in a fast and convenient way eradicating the elimination of the necessity for time-consuming manual searches using keywords.

• Deeper understanding of documents: Get more than literal matches with the keywords present in the source texts. thus, it escalates to an enhanced discovery of-knowledge.

• Increased productivity: They reduced the time and energy that people used to spend in searching methods that are conventional research and analysis.

IV.METHODOLOGY OF PROPOSED SURVEY

RAG (Retrieval Augmented Generation) Model:

RAG is a method for integrating new information into LLM model prompt and placing it into the model [7].

A typical RAG application has some main components:

- **Load:** To begin, we should load data into memory. It can be done with Document Loaders.
- **Split:** Text splitters as the name suggests divides Documents into a number of parts. The chunking is useful both at the index time and at the model input time.
- **Indexing:** It is a system whereby data is acquired from a source and made ready for indexing. This usually happens offline.
- **Retrieval and generation:** the actual RAG chain that happens at run time and provides the mechanism that given a user query, fetches the required data from the index and then feeds it to the model.
- **Store:** It then requires space to save and archive them so that they can be searched.

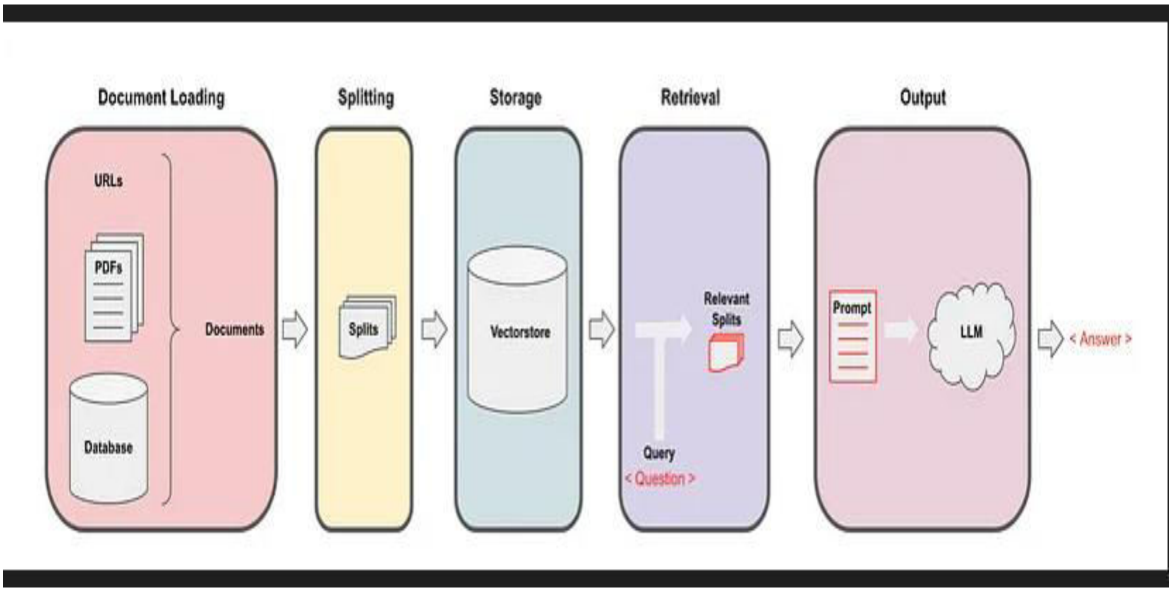

Figure 1 : RAG Architecture

LLM (Large Language Models)

GPT-4 is an advanced version of language modeling that involves understanding the textual content, its context, and the generation of text that is probably going to be written by a human. That's why, learning a variety of datasets, it can work on the tasks like creation of questions, analysis of text, and summarization.

Content extraction takes form through identifying important points and then generalizing an idea from the book, then summarizing each chapter.

Question Generation Based on the content extracted, develop application, comprehension and analytical questions [8].

LANG CHAIN

Lang chain is an advanced, open-source framework, is made specifically to help you in creating applications that use language models, primarily large language models (LLMs). The fundamental concept of the library is that various parts can be "chained" together and generate progressively complex use cases based around LLMs. Lang Chain is made up of many parts from various modules [7].

Prompts: Using templates, this module lets you create dynamic prompts. Depending on the input variables and the context window's size (conversation history, search results, prior replies, and more) that are utilized as context, it can adjust to various LLM kinds.

Models: To connect to the majority of third-party LLM APIs, this module offers an abstraction layer. It supports conversation, embedding models, and offers API links to about forty public LLMs.

FAISS (Facebook AI Similarity Search)

A library called FAISS (Facebook AI Similarity Search) is used to efficiently search for similarities and cluster dense vectors. It can search multimedia materials (such as photos) in ways that ordinary database engines (SQL) cannot or will not do. Algorithms within it can search through sets of collections of vectors, regardless of size, even those that might not fit in RAM. It also includes accompanying code for parameter adjustment and evaluation[9].

V.CONCLUSION AND FUTURE WORK

Imagine exploring several PDFs with ease, avoiding time-consuming keyword searches, and conversing naturally instead. This ambition is realized through the "Document Based Question Answering System using Python and Lang Chain" initiative, which makes use of sophisticated language models and blazingly quick vector storage. By using FAISS, quick vector storage, and LLM and Lang Chain, a complex language model, uploaded documents may be easily interacted with using simple English queries.

This will be capable of generating analytical queries along with their corresponding answers. We will make it more user-friendly and efficient to have it respond on its own to e-commerce clients' inquiries very fast [9].

In future work, we intend to extend the model methodology that we have proposed and study the recent emergence of more advanced LLMs in this fast-changing area of generative language models. Given the rapid growth in GenAI, our research activities will be oriented toward studying, integrating, and comparing with the very newest developments in this sector. Moreover, our empirical study provides evidence of a relative bias in the conceptual questions generated based on a given test document and a decrease in the quality of generated questions during test cycles. We will like to visualize that using explainable AI, XAI, downstream layer representation techniques. Future studies also aim at broadening our study agenda to further the multilingual and multi-file document model interpretability for general performance, inclusivity, and efficacy.

VI.RESULTS

LLM has successfully learned to produce suitable extracts, summaries and or questions for the texts extracted and pre-processed from the PDFs. LangChain helps with keeping the workflow synchronized, and FAISS is efficient and performs the search quickly. The outcomes prove the effectiveness and the possibility of the presented approach for affective enriching of the analysis of the educational content and the assessment of the knowledge.

REFERENCES

1. Smith, J., Brown, A., & Johnson, C. A. “Obstacles Identified in Manual Reading for Information Extraction from PDF Documents”, The Ninth International Conference on Natural Language Processing : Proceedings-2024.
2. Automatic Question Generation Using Software Agents for Technical Institutions by Shivank Pandey1, K.C. Rajeswari International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-4 Issue-13 December-2020
3. A.S.M Mehedi Hasan , Md. Alvee Ehsan, Kefaya Benta Shahnoor “Automatic Question & Answer Generation Using Generative Large Language Model (LLM)”. vol. 22, pp. 457–479, 2024.
4. M. Bidoki, M. Fakhrahmad, and M. Moosavi, “Text summarization as a multiobjective optimization task: Applying harmony search to extractive multidocument summarization,” The Computer Journal, vol. 65, no. 5, pp. 1053– 1072, 2022.
5. Jones, R., & Patel, S. “Challenges Associated with the Use of Keywords in PDF Files”. Journal of Information Retrieval, 45(2), 123-135.

6. G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support by Ming Liu and Rafael A. vol. 40, no. 14, pp. 5755–5764, 2019.
7. Devlin, J., Cary, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. [cs.AI] arXiv:2305.07871v1 May 13, 2023
8. Bidyut Das, Mukta Majumder, and Santanu Phadikar's survey, which included automated question development and response evaluation Das, Arif Ahmed Sekh et al. Technology-Enhanced Learning Research and Practice, 2021, 16:5.
9. Automatic Multiple Choice Question Generation System for Semantic Attributes Using String Similarity Measures by Ibrahim Eldesoky Fattoh in journal Computer Engineering and Intelligent Systems www.iiste.org ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.5, No.8, 2014.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com