DOI: http://dx.doi.org/10.26483/ijarcs.v15i2.7064

ISSN No. 0976-5697

Volume 15, No. 2, March-April 2024

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CLOSED-DOMAIN QUESTION-ANSWERING SYSTEM FOR INDIAN LEGAL ACTS

Ankita Yadav
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India.

Prof. Piyush Pratap Singh
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India

Abstract: In Today's digital age, access to legal knowledge among students is crucial for fostering a well-informed citizenry. However, comprehending the intricacies of provisions of Indian Legal Acts, especially in a complex legal system like India's, can be daunting, even for seasoned professionals, let alone for educational purposes among children. To bridge this gap and foster legal literacy from a young age, we introduced a PDF-based Closed-domain Question-Answering (CDQA) System for Indian Legal Acts. Our system simplifies educating students about provisions of Indian Legal Acts such as RTI(Right to Education), RTE(Right to Education), Anti-Dowry Act, etc. By leveraging natural language processing(NLP) techniques and machine learning algorithms, our system, powered by LangChain, enables users to pose questions in natural language related to Indian Legal Acts. LangChain represents a groundbreaking advancement in question-answer systems, harnessing the power of cutting-edge language models to provide accurate and comprehensive responses to user queries. Developed based on state-of-the-art NLP techniques, LangChain is a versatile and highly adaptable tool that offers a versatile solution with transformative potential across diverse industries, including education, healthcare, legal research, customer support, etc.

Keywords: Indian Legal Acts, CDQA, LangChain, PDF-based Question-Answering System, RTI ,RTE etc

I. INTRODUCTION

Question-answering systems have become quite popular for retrieving relevant information in a short time. In today's digital era, many structured and unstructured data are generated daily through news articles, social media platforms, websites, etc. Extracting relevant and precise information from that bulk data is daunting [1]. Question-answering systems bridge the gap by enabling users to query structured datasets effectively, facilitating informed decision-making and comprehensive information retrieval. It needs a query from the user side that is further analyzed by the question-answering system, and then the relevant answer is fetched from the corpus[2]. Closed-domain Question -Answering Systems are domain-specific. Hence, their accuracy is high[3]. It is designed to respond to queries within a specific domain, such as Indian legal acts. For example, If someone wants to know, "Which age group does the Right to Education (RTE) act apply to?" then instead of going through multiple articles on the internet, it can provide a direct and accurate answer to the question, such as the age group to which the RTE act applies is 6 to 14 years old.

LangChain introduces a groundbreaking approach to question-answering systems, leveraging the wealth of information in PDF documents. With its innovative framework, LangChain seamlessly integrates advanced Natural Language Processing techniques with the structured format of PDFs, enabling efficient extraction and comprehension of textual data[4]. By harnessing the power of Large Language models and sophisticated algorithms, LangChain empowers users to pose queries in natural language and receive accurate and comprehensive answers extracted directly from PDF documents. This PDF-based question-answering system revolutionizes information retrieval, facilitating rapid access to knowledge encapsulated within legal documents, educational materials, and more. Whether unraveling the complexities of legal acts or educational content, LangChain sets a new standard for efficiency and accessibility in navigating the vast sea of information stored in PDF. LangChain operates through a sophisticated mechanism driven by a large language model (LLM) such as GPT (Generative Pre-Trained Transformer), augmented by prompts, chains, memory management, and question-answer retrieval [5]. The LLM forms the core, providing linguistic understanding. Prompts structure user interactions, while chains dictate system flow. Memory stores and retrieves contextual information, enhancing response accuracy. Question-answer retrieval matches user queries to store data, delivering relevant answers [6]. This mechanism harmoniously integrates advanced NLP techniques, facilitating efficient natural language processing and information retrieval tasks adaptable to diverse domains [7].

In this paper, the section 'II' presents the Literature Survey. The section 'III' is focused on data collection and methodology. Section 'IV' will demonstrate the result and discussions of the question-answering system. At last, we will proceed to the conclusion in section 'V.'

This research aims to build up a PDF document-based closed-domain question-answering system for Indian legal acts for educational purposes so that students can get complex information related to these acts simply and more easily.

II. LITERATURE SURVEY

Information retrieval is a key component within Natural Language Processing. It aims to extract relevant answers from a given passage in response to a passed question[8]. In contemporary society, information retrieval is crucial in everyday activities, serving various practical purposes. Its applications span web searches, question-answering systems, personal assistants, chatbots, and among other areas. The core objective of information retrieval is to efficiently locate and retrieve information that matches a user's particular query [9]. Various model types are utilized in large language models(LLMs), such as those employed in LangChain. Predominantly, LLMs are central to the framework, operating by receiving a text string(prompt) and generating a corresponding text string as output. Additionally, LangChain incorporates other model types, including chat models and text embedding models. Chat models are tailored explicitly with structured APIs to process chat messages efficiently. On the other hand, Text Embedding models are adept at converting text inputs into corresponding embeddings, represented as lists of floating-point numbers. These embeddings are crucial when dealing with custom documents within the LangChain system [5]. LangChain offers a comprehensive set of functionalities to streamline the loading, transformation(splitting), storage, and querying of data. It provides the following classes to facilitate these operations [5]:

- Document Loaders: These classes can load documents from various sources, including CSV, PDF, HTML, JSON, Excel, GitHub, Google Drive, etc.
- Document Transformers: Designed to split documents into smaller chunks, enabling efficient processing by large language models (LLMs).
- Text Embedding Models: These models input unstructured texts and generate a list of floating-point numbers representing corresponding embeddings.
- Vector Stores: These classes aid in storing and retrieving the embedded data, facilitating efficient search operations.
- Retrievers: Utilized to query data based on similarity metrics derived from embeddings.

In LangChain, embeddings numerically represent text, aiding similarity search assessment and input selection for language models. Stored in vector databases, they enable efficient retrieval of semantically similar text. User queries, converted to embeddings, are compared with database entries using dot products or cosine similarity. The similarity score,

$$\mathbf{a} * \mathbf{b} = |\mathbf{a}| * |\mathbf{b}| * \mathbf{cos}(\alpha),$$

determines relevance based on semantic proximity. Where a and b are the vectors being compared, ' |a|' and '|b|' are the magnitudes of the vectors, and '**cos(**α)' is the cosine of the angle between them[10].

III. DATA AND METHODOLOGY

A. *LangChain Framework*

LangChain is a comprehensive framework for developing a question-answering (QA) system, developed by OpenAI, that utilizes advanced natural language processing (NLP) techniques and large language models to provide accurate and comprehensive responses to user queries. Here's an explanation of the key aspects of the LangChain mode:

- Language Model: LangChain is built upon a powerful language model developed by OpenAI, typically based on transformer architecture, such as GPT(Generative Pretrained Transformer). This language model undergoes extreme pre-training on vast amounts of text data, enabling it to understand and generate human-like text across various domains and topics. Through pretraining, the language model acquires the ability to capture semantic, syntactic, and contextual information from textual data, facilitating accurate and fluent responses.

- Question-Answering with LangChain: LangChain specializes in question-answering tasks, allowing users to input natural language queries into the system. Leveraging its understanding of the query and the underlying text data, LangChain generates relevant answers with remarkable accuracy and fluency. This process involves mechanisms such as attention, enabling the model to focus on pertinent parts of the input text when formulating the process.
- Fine-tuning for Domain-Specific Tasks: LangChain undergoes a fine-tuning process to optimize its performance further. The base language model is trained on specific domains or datasets relevant to the intended application during fine-tuning. This adaptation process enables LangChain to tailor its language understanding capabilities to particular tasks or domains, resulting in improved accuracy and relevance of responses [11].

- Integration of Information Retrieval Techniques: In addition to language modeling and question answering, LangChain incorporates information retrieval techniques to search through large collections of text data efficiently. This may involve indexing and storing text data in a structured format, facilitating fast retrieval of relevant documents or passages based on their similarity to a given query.

B. *Architecture of the Methodology*

Here are the steps of the methodology:

a) **Compiling a PDF for Indian Legal Acts**: I have compiled a PDF document containing details of 84 Indian Legal Acts sourced from diverse resources, including government websites like "India code.nic.in" and "india.gov.in" portal, along with information from NCERT books and some general knowledge websites

.

b) **PDF Processing**: PyPDF2 library is used to read and extract text from PDF documents containing information regarding provisions of Indian Legal Acts.

c) **Text Splitting**: The text is split into smaller chunks using a character-based text splitter, allowing for efficient indexing and retrieval.

d) **Embedding Generation**: OpenAI embeddings are generated for the text chunks. These embeddings capture semantic and contextual information about the text, facilitating similarity computation and retrieval.

e) **Vector Storage**: FAISS (Facebook AI Similarity Search) is used as a vector store to store and retrieve embeddings efficiently. FAISS employs index structures and algorithms optimized for fast similarity search in high-dimensional spaces.

f) **Question- Answering:** The code utilizes LangChain's QA capabilities, which involve processing user queries and retrieving relevant documents or passages from the indexed text data. The LangChain model employs attention mechanisms and transformer-based architectures to understand the context and semantics of queries and text data. The system uses similarity search algorithms to match user queries with the most relevant documents or passages based on the embeddings and their similarities.

g) **Retrieval-based Question -Answering**: In addition to traditional QA, the script demonstrates a retrieval-based approach, where relevant documents are retrieved based on similarity to the query using FAISS. Then, LangChain processes the retrieved documents to provide an answer to the query.

h) **Language Model**: LangChain leverages OpenAI's language model, a GPT (Generative Pre-trained Transformer) architecture variant. GPT models are based on transformer neural networks and pre-trained on large corpora of text data to understand and generate human-like text.

i) **Answer Display**: At the end, answer of the query is displayed.

Overall, the combination of PDF processing, text splitting, language modeling, embedding generation, and retrieval-based techniques enables the LangChain system to answer questions related to Indian Legal Acts effectively.

➢ What is FAISS?

FAISS (Facebook AI Similarity Search) is an efficient library for similarity search and clustering of large-scale datasets of vectors. It is primarily used for nearest neighbor search, which involves finding the vectors in datasets most similar to a given query vector. FAISS is particularly well-suited for high-dimensional vector spaces. It is an essential tool in many machine learning and natural language processing tasks, including information retrieval, recommendation systems, multimedia search, natural language processing, and anomaly detection.

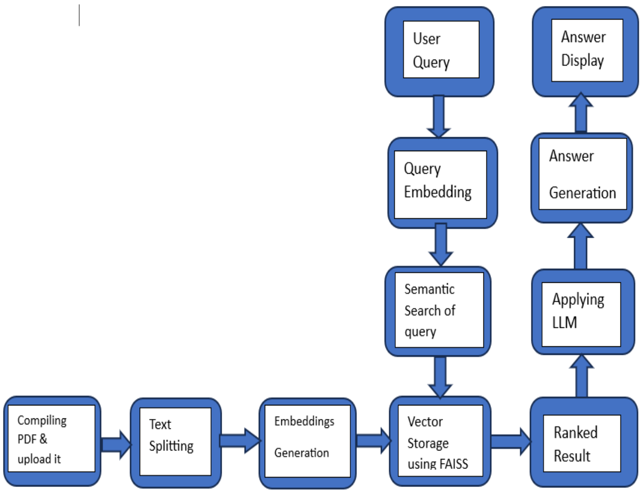

Fig 1 The architecture of the proposed methodology

IV. RESULT

In figure no. 2, The code first loads a Question-Answer chain using the 'load_qa_chain' function with OpenAI as the language model and 'map_rerank' as the chain type. It also specifies 'return_intermediate_steps=True' to retrieve intermediate steps during the question-answer process. It then retrieves a set of documents('docs') using similarity search based on the query "What are the provisions of RTE?" from the 'docsearch' object. The QA chain is then applied to the retrieved documents and the query, and the'results' variable contains the output of the QA chain. Upon processing the query, the LangChain-powered system presents a concise yet comprehensive paragraph summarizing the key provisions outlined within the document. Each provision is meticulously detailed, covering aspects such as Right to Education, government obligations, financial considerations and other factors delineated within the PDF. The paragraph not only address the specific query posed by the user but also offers additional context and insights derived from the document's content.


```
chain = load_qa_chain(OpenAI(),
                      chain_type="map_rerank",
                      return_intermediate_steps=True
                      )

query = "what are the provisions of RTE?"
docs = docsearch.similarity_search(query,k=15)
results = chain({"input_documents": docs, "question": query}, return_only_outputs=True)
results
```

```
{'intermediate_steps': [{'answer': ' The provisions of RTE include the right to free and compulsory education until completion of elementary education, the obligation of the government to ensure admission, attendance, and completion of education, free education for children, admission of non-admitted children, duties of governments and parents, and sharing of financial burden between central and state governments.',
   'score': '100'},
  {'answer': ' The provisions of the RTE Act are briefly described below. The Act provides for:',
   'score': '100'},
  {'answer': ' The RTE Act mandates for all private schools to reserve 25 per cent of their seats for children from socially disadvantaged and economically backward sections. This provision is included in Section 12(1)(c) of the RTE Act. All schools (private, unaided, aided or special category) must reserve 25% of their seats at',
   'score': '100'},
```

Fig 2 Sample Question-Answer regarding RTE

Furthermore, in figure no.3, the LangChain's model's ability to assign scores to intermediate steps provides transparency into the reasoning behind the selected provisions, allowing users to understand the rationale for each included detail. This ensures that users can trust the accuracy and relevance of the system's information.

```
[ ] results['intermediate_steps']

[{'answer': " The provisions of the RTE Act include the right to free and compulsory education for children until they complete their elementary education, the obligation of the government to ensure admission, attendance, and completion of education for children between the ages of six and fourteen, the admission of non-admitted children to their appropriate age class, and the duties of governments, local authorities, and parents in ensuring a child's education. The Act also specifies the sharing of financial burden between the central and state governments.",
  'score': '100'},
 {'answer': ' The provisions of the RTE Act are briefly described below. The Act provides for:',
  'score': '80'},
 {'answer': ' The RTE Act mandates for all private schools to reserve 25 per cent of their seats for children from socially disadvantaged and economically backward sections. This provision is included in Section 12(1)(c) of the RTE Act. All schools (private, unaided, aided or special category) must reserve 25% of their seats at',
  'score': '80'},
 {'answer': ' The Act lays down specific standards for the student-teacher ratio, which is a very important concept in providing quality education. It also talks about providing separate toilet facilities for girls and boys. ',
  'score': '80'},
 {'answer': ' Children below 6 years are not covered under the Act.',
  'score': '80'},
 {'answer': ' This provision is included in Section 12(1)(c) of the RTE Act. All schools (private, unaided, aided or special category) must reserve 25% of their seats at the entry level for students from the Economically Weaker Sections (EWS) and disadvantaged groups. This provision is a far-reaching move and perhaps the most important step in so far as inclusive education is concerned. This provision seeks to achieve social integration. The loss incurred by the schools as a result of this would be reimbursed by the central government. The Act has increased enrolment in the upper primary level (Class 6-8) between 2009 and 2016 by 19.4%.'
```

Fig 3 Sample Question-Answer regarding RTE

```
[25] query = "when did right to education act come into force ?"
     docs = docsearch.similarity_search(query_02)
     chain.run(input_documents=docs, question=query)

'\n\nThe Right to Education Act came into force in 2010.'

[29] query = "what is article 21A?"
     docs = docsearch.similarity_search(query_02)
     chain.run(input_documents=docs, question=query)

' Article 21A is an important constitutional amendment that was inserted in the Indian Constitution in 2002, which states that the State shall provide free and compulsory education to all children between the ages of 6-14. This amendment marks the right to education as a fundamental right in the country.'
```

Fig 4 Sample Question -Answer regarding RTE

In Figure 4, the code snippet executes a query, "When did the Right to Education Act come into force?" and then runs a QA chain to generate an answer. The generated answer provides the specific year when the Right to Education Act was enacted in 2010. Again, it poses a second query, "What is article 21A?" and runs a QA chain to generate an answer. The generated answer summarizes Article 21A of the Indian Constitution, highlighting its free and compulsory education provision for children aged 6 to 14.

In Fig 5, we can see a sample paragraph of the PDF containing information regarding the Right to Education Act. The system can get the answer to the question from the PDF document. We can see that the response matches the PDF document. By calculating the map-rerank score of each query, we can find the accuracy of any answer. (as shown in Fig 3 and Fig 4).

Right to Education Act

The Act is completely titled **"the Right of Children to Free and Compulsory Education Act"**. It was passed by the Parliament in August 2009. When the Act came into force in 2010, India became one among 135 countries where education is a fundamental right of every child.

- The 86th Constitutional Amendment (2002) inserted Article 21A in the Indian Constitution which states:
  - "The State shall provide **free and compulsory education to all children of 6 to 14** years in such manner as the State, may by law determine."
- As per this, the right to education was made a fundamental right and removed from the list of Directive Principles of State Policy.

The RTE is the consequential legislation envisaged under the 86th Amendment **RTE Provisions**

The provisions of the RTE Act are briefly described below. The Act provides for:

- The right of free and compulsory education to children until they complete their elementary education in a school in the neighbourhood.
- The Act makes it clear that 'compulsory education' implies that it is an obligation on the part of the government to ensure the admission, attendance and completion of elementary education of children between the ages of six and fourteen. The word 'free' indicates that no charge is payable by the child which may prevent him/her from completing such education.
- The Act provides for the admission of a non-admitted child to a class of his/her appropriate age.
- It mentions the duties of the respective governments, the local authorities and parents in ensuring the education of a child. It also specifies the sharing of the financial burden between the central and the

Fig 5 Sample paragraph from PDF

V. CONCLUSION

The result from the PDF-based question-answering system, powered by the LangChain model, provides valuable insights into the provisions outlined within the document related to the query. The LangChain model leverages state-of-the-art natural language processing techniques to analyze the content of the PDF comprehensively, extracting pertinent information that directly addresses the user's query. Through sophisticated techniques such as mapping and reranking, integrated seamlessly within the LangChain framework, the system identifies passages with high relevance and confidence levels, ensuring the accuracy and reliability of the provided answer.

This system will help enhance students' knowledge regarding Indian Legal Acts. We can also update our PDF document by collecting information on new acts in the future.

VI. REFERENCES

[1] R. Mervin, "An Overview of Question Answering System" International Journal Of Research In Advance Technology In Engineering (IJRATE) Volume 1, Special Issue, October 2013 Proceedings of National Conference on Recent Trends In Web Technologies- Rtwt 2013.

[2] Shreya Acharya, K. Sornalakshmi*, Bidisha Paul, Anshul Singh,2022, "Question Answering System using NLP & BERT" Proceedings of the Third International Conference on Smart Electronics and Communication (ICOSEC 2022) IEEE Xplore Part Number: CFP22V90-ART; ISBN: 978-1-6654-9764-0.

[3] MAHSA ABAZARI KIA JON CHAMBERLAIN, AND SHOAIB JAMEEL, AYGUL GARIFULLINA, (Member, IEEE), MATHIAS KERN "Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion" date of publication April 25, 2022, date of current version May 3, 2022. Digital Object Identifier 10.1109/ACCESS.2022.3170466.

[4] Rakha Asyrof, Muhammad Irfan Lutfhi, Mutia Rahmi Dewi, Prasetyo Wibowo, "Systematic Literature Review Langchain Proposed" 2023 International Electronics Symposium (IES) | 979-8-3503-1473-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/IES59143.2023.10242497.

[5] Oguzhan Topsaka, T. Cetin Akinc "Creating Large Language Model Applications Utilizing LangChain: A sPrimer on Developing LLM Apps Fast" Article in International Conference on Applied Engineering and Natural Sciences · July 2023 DOI: 10.59287/icaens.1127.

[6] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, "PromptChainer: Chaining Large Language Model Prompts through Visual Programming" arXiv:2203.06566v1 [cs.HC] 13 Mar 2022.

[7] Keivalya Pandya, Prof. Dr. Mehfuza Holia "Automating Customer Service using LangChain Building custom open-source GPT Chatbot for organizations", 3rd International Conference on "Women in Science & Technology: Creating Sustainable Career" 28 -30 December, 2023.

[8] . K. Lanyo and A. Wausi, "A Comparative Study of Supervised and Unsupervised Classifiers Utilizing Extractive Text Summarization Techniques to Support Automated Customer Query Question-Answering," in

2018 5th International Conference on Soft Computing Machine Intelligence (ISCMI), Nov. 2018, pp. 88–92. doi: 10.1109/ISCMI.2018.8703237.

[9] . I. Thalib, Widyawan, and I. Soesanti, "A Review on Question Analysis, Document Retrieval and Answer Extraction Method in Question Answering System," in 2020 International Conference on Smart Technology and Applications (ICoSTA), Feb. 2020, pp. 1–5. doi: 10.1109/ICoSTA48221.2020.1570614175.

[10] LangChain Use Case Examples, https://docs.langchain.com/docs/category/use-cases Accessed July 10th, 2023.

[11] Humza Naveed, Asad Ullah Khan, Shi Qiu2, Muhammad Saqib, Saeed Anwar5, Muhammad Usman, Naveed Akhtar, Nick Barnes, Ajmal Mian. " A Comprehensive Overview of Large Language Models" arXiv:2307.06435v8 [cs.CL] 20 Feb 2024.