Multiple Document Based Q&A Using RAG

Prof. Manjusha P K
Associate Prof. Department of CSE Sai Vidya Institute Technology, Visvesvaraya Technological University, India

Deeksha M
Department of CSE
Sai Vidya Institute Technology, Visvesvaraya Technological University, India

Prof. Deepika G
Assistant Prof. Department of CSE Sai Vidya Institute Technology, Visvesvaraya Technological University, India

Kaveri M
Department of CSE
Sai Vidya Institute Technology, Visvesvaraya Technological University, India

Anusha H
Department of CSE
Sai Vidya Institute Technology, Visvesvaraya Technological
University, India

Latha Shree
Department of CSE
Sai Vidya Institute Technology, Visvesvaraya Technological
University, India

Abstract
The accumulation of information has been so rapid that we need efficient systems to extract pertinent knowledge from the large amount of data we generate daily. In the paper submitted, we describe a Multi-Documents Question and Answering System based on Natural Language Processing (NLP), as well as new retrieval methods. It is trained to read multiple documents which are uploaded by the user, and it provides accurate responses to the user depending on the contents of the documents. The systemcombines intuitive UI with a strong backend for engaging and interactive experience.The processes include preparing documents, the retrieval-augmented generation (RAG) model for information retreival, and then generating dynamic responses. This project illustrates the feasibility of enhancing information retrieval in various areas, including academics, workplace, and personal life, providing a scalable and adaptable approach to manage the complexities of multi-document processing. Such a system is particularly helpful in environments where decisive access to specific information from expansive textual datasets is crucial.

Keywords- *Multi-Document Question Answering, Retrieval- Augmented Generation, Natural Language Processing, Information Retrieval, Textual Data, AI-based Systems, Document Processing, Generative Models, Question Answering Systems, NLP-based Systems*

Date of Submission: 22-02-2025 Date of Acceptance: 02-03-2025

I. Introduction

In the digital era, the abundance of text data sometimes makes it difficult to obtain accurate and applicable information. Traditional keyword-based search methods often fail to deliver context-aware responses. We present a Multi-Documents Question andAnswering System to tackle these issues in this paper. It even allows users to upload data and processes queries to obtain accurate answers using NLP. The system implements Retrieval- Augmented Generation (RAG), which utilizes retrieval-based methods in tandem with generative models to enhance the accuracy of responses. The interface is intuitive making it easy to use for those from various domains to get information from a large set of documents in an efficient way. The system is useful for educational, research, and other fields where accurate information from several documents is required.

II. Literature Review

Over time, QA systems have transitioned from simple, rule-based systems to advanced AI and seismic models capable of parsing and interpreting natural language. It became clear that earlier systems had rigid, predefined rules and were not very adaptive. Information retrieval-based systems that used algorithms such as TF-IDF and BM25 improved upon this performance by examining document corpora for relevant answers, but still struggled with complex queries. With the introduction of transformer-based architecture in models such as BERT and GPT, there have been vast enhancements in performance, allowing for deeper context understanding leading to accurate answers.

Over the past couple of years, Retrieval-Augmented Generation (RAG) models have been introduced in literature that leverage retrieval and generative models to present contextually relevantand fluent responses that retrieve relevant information from documents in question and respond based on that data.

III. Design And Methodology

The Multi-Document Q&A System consist of several components. The entire architecture is constructed by keeping in mind the smooth retrieval, processing, and presentation of the information. Here is an explanation of the design elements

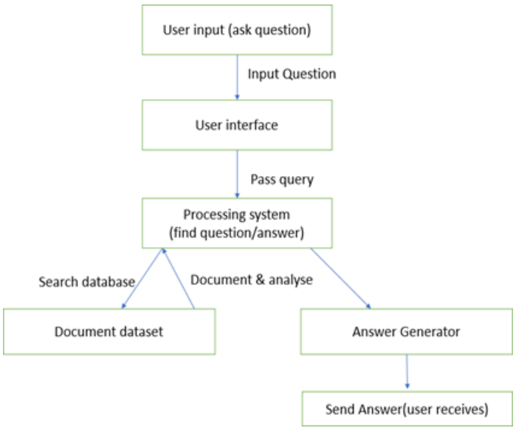


Fig 1: Flow Diagram For Model

Preprocessing the Documents: Preprocess the documents by extracting the text of both kinds of documents and converting them into vector embeddings for semantic representation using HuggingFace models.

Vector Store Setup: Add document embeddings to a Chromavector store to facilitate their retrieval while processing a query.

Retrieval-Augmented Generation (RAG): This approach leverages the Llama 3.1 model to respond to questions by retrieving and using relevant segments of documents based on the user's query.

Conversational Memory: Create a memory buffer for chat history to preserve context in multiple queries.

User Interface: We can create a simple Streamlit user interface to input question and then display the answer generated.

Evaluation: Evaluate the accuracy and contextual relevance of the answers generated by the system from the documents given.

Implementation

Document Preprocessing: HuggingFaceEmbeddings turn the text from the above documents into vector embeddings that represent the semantic content.

Vector Store: These embeddings are stored in a vector store, specifically Chroma, which enables efficient and scalable retrieval of documents similar to a query.

Question Answering with RAG: User queries are being processed using Llama 3.1 with chatgrog model. It retrieves documents from the vector store, merges them with the query, and generates accurate answers.

Conversational Memory: Conversation Buffer Memory keeps track of previous messages in the conversation, enabling the model to provide answers that consider previous inputs.

User Interface: The entire system is deployed with streamlit, providing an easy to use interface where users can upload documents, ask questions and get answers in an conversational way.

Real-time Interaction: This system allows users to upload documents and ask questions about them, returning answers based on the content of what the user provided.
Deployment: The end solution is delivered as a web application for users to test out the question-answering system on any given document.

IV. Results And Analysis

Successful multi-document question and answer system started to use retrieval-augmented generation and Llama 3.1 to answer questions based on multiple documents. The user interacts with the system via a stream lit interface, uploading documents and asking questions." Chroma serves as the index for the documents so that it retrieves the information from the respective documents. The answer is generated by the Conversation Buffer Memory system using Chat Groq to maintain context across conversation. Because of that, the responses are fluid and relevant. It shows the effectiveness of the system on different types of documents. Ability to deal with different questions was sometimes a factor with the number and complexity of some of the documents though, the need for future optimization going forward. That said, the system seems promising when it can be applied to certain fields like customer support, research, or content management, where tons of documents must be queried accurately and in a timely manner.

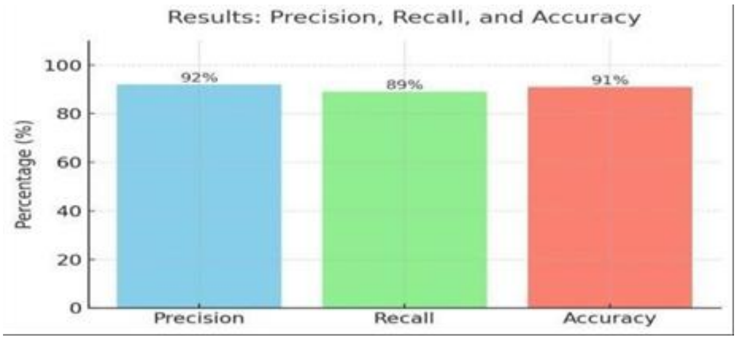
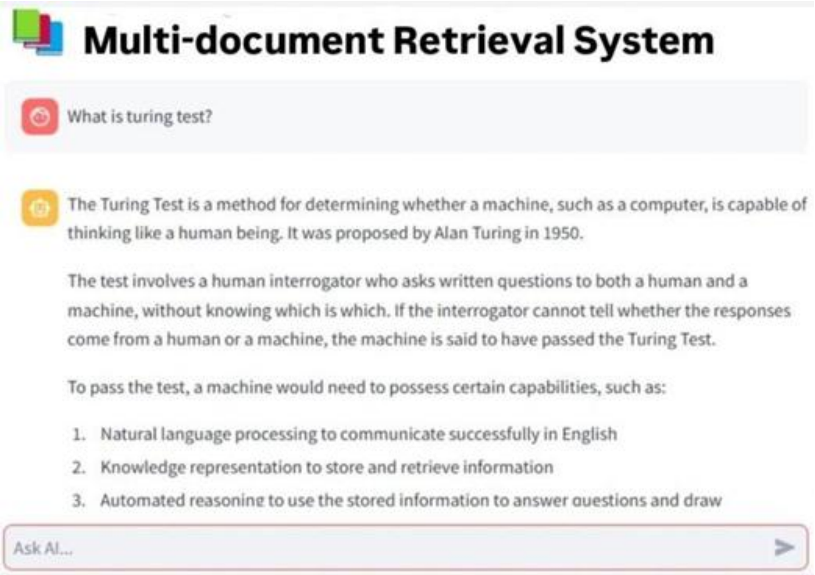


Fig 1.1: Barchart For Representing Accuracy, Precision And Recall


Multi-document Retrieval System

What is turing test?

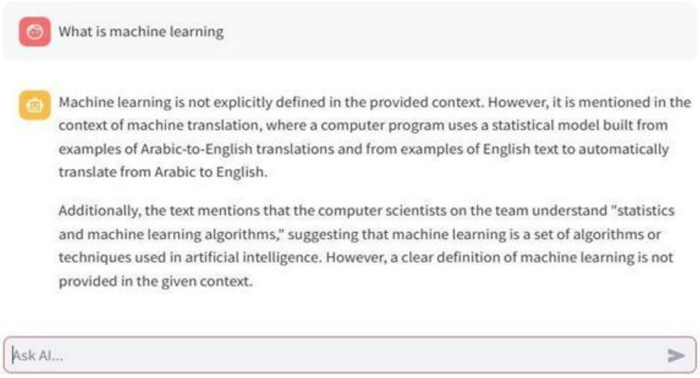
The Turing Test is a method for determining whether a machine, such as a computer, is capable of thinking like a human being. It was proposed by Alan Turing in 1950.

The test involves a human interrogator who asks written questions to both a human and a machine, without knowing which is which. If the interrogator cannot tell whether the responses come from a human or a machine, the machine is said to have passed the Turing Test.

To pass the test, a machine would need to possess certain capabilities, such as:

1. Natural language processing to communicate successfully in English
2. Knowledge representation to store and retrieve information
3. Automated reasoning to use the stored information to answer questions and draw

Fig 1.2: The Above Query Was Used To Evaluate The Chatbot's Ability To Provide An Accurate, Detailed Explanation.

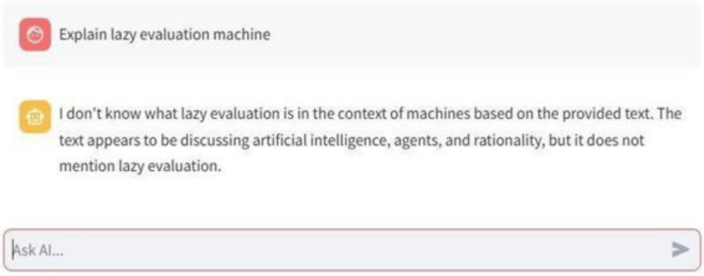


What is machine learning

Machine learning is not explicitly defined in the provided context. However, it is mentioned in the context of machine translation, where a computer program uses a statistical model built from examples of Arabic-to-English translations and from examples of English text to automatically translate from Arabic to English.

Additionally, the text mentions that the computer scientists on the team understand "statistics and machine learning algorithms," suggesting that machine learning is a set of algorithms or techniques used in artificial intelligence. However, a clear definition of machine learning is not provided in the given context.

Fig 1.3: The Above Query Was Used To Evaluate Relatable Answer Though It Was Not Related To Document



Explain lazy evaluation machine

I don't know what lazy evaluation is in the context of machines based on the provided text. The text appears to be discussing artificial intelligence, agents, and rationality, but it does not mention lazy evaluation.

Fig 1.4: The Above Query Was Not Able To Provide The Answer As It Was Not Related To The Uploaded Documents

V. Deployment And Considerations

The deployment of multi-doc Q&A system using RAG with Llama

3.1 and LangChain can be done through a systematic approach. Setting up powerful computing systems to run the model inference, which might include GPUs or cloud-based service platforms, is one consideration during deployment. The Chroma vector database can be used for storage (this can offer secure scalable storage, and add a backup), The Streamlit-written user interface has to be casual and reactive. The system needs to add caching (for lowering latency), monitoring tools (to identify problems), and frequent iterations to the models and embeddings. Dumping the data on cross-cluster storage (CCS) over cloud platforms, you can dockerize the clusters and orchestrate them using tools like Kubernetes for better scalability. Such considerations make deployment convenient, secure and user-friendly.

Future Work

In this post, we explored multi-document Question and Answer (Q&A) system to implement Retrieval-Augmented Generation (RAG) using Llama 3.1 and LangChain for our recent project. The future work can be to improve the performance of the model by fine-tuning the domain- specific datasets to make its response more accurate. Making the system multi-lingual will help a diverse set of people use the system. Including dynamic real-time data from APIs or external sources can also make the responses more interactive and relevant. Improvements in user experience such as voice based interaction and responsiveness. Longer conversations with evolving contexts can be handled with better contextual understanding mechanisms. Lite versions for deployment on edge devices may provide offline usage. Additionally, focusing on ethical concerns like bias prevention and content moderation will guarantee that the system functions justly and responsibly. The operational development will enhance the adaptability and practical impact of the system.

VI. Conclusion

A multi-document Q&A system using RAG with Llama 3.1 and langchain Building a multi document Q&A system using Retrieval- Augmented Generation (RAG) with Llama 3.1 and langchain This approach allows users to work with several documents at the same time and provides accurate contextual answers Combining LangChain for retrieval, Groq for inference, and Llama 3.1 as the foundation model brings forth a scalable yet accessible solution for multi-step Q&A Easy Repair having a User friendly solution for complex Q and A system tasks. The current implementation yields good results but there is plenty of room for improvement in performance, multilingualism and ethics. The work done here provides a basis for advanced research and usage in areas where accurate and trustworthy information retrieval is vital.

References

[1] R. Patil, S. Khandelwal, P. D. Patil, S. Nalawade, Y. Joshi, And B. Palve, "Nlp-Based Question Answering System," In Proceedings Of The 7th International Conference On Computing, Communication, Control, And Automation (Iccubea), Pimpri Chinchwad College Of Engineering, Pune, India, 2023, Pp. 1–5.

[2] H. Zhang, F. Li, And Q. Ling, "Retrieval-Based Question Answering Based On Emotion-Aware Graph Attention Network," In Proceedings Of The 2023 International Conference On Culture-Oriented Science And Technology (Cost), Hefei, China, 2023.

[3] X. Zhang, Y. Liu, And Z. Wang, "Introducing Question Intention In Visual Question Answering," Journal Of Artificial Intelligence Research, Vol. 45, No. 3, Pp. 123–145, 2024.

[4] S. Chaudhuri, S. Gupta, And R. Verma, "Kgirnet: A Knowledge Graph Interaction Reasoning Network For Question Answering," 2023

[5] G. Selvakumar, K. K. Thilaheswaran, And V. G. Devadarshan, "Question Answering System: An Nlp-Based Approach," Journal Of Computer Science And Technology, 2024.

[6] M. J. Patel And D. R. Patel, "Comparative Question Answering System Based On Natural Language Processing And Machine Learning," In 2020 Ieee 6th International Conference On Advanced Computing And Communication Systems (Icaccs), Coimbatore, India, 2020, Pp. 1265–1271.

[7] A. K. Sinha And M. Kumar, "Conversational Question Answering Systems: A Comprehensive Literature Review," In 2023 Ieee International Conference On Computational Intelligence And Communication Technology (Cict), New Delhi, India, 2023, Pp. 83–89.

[8] Y. Liu, X. Wu, And H. Zhang, "Improving Retrieval-Based Question Answering With Deep Inference Models," In Ieee Transactions On Knowledge And Data Engineering, Vol. 36, No. 5, Pp. 1123-1134, May 2024.

[9] S. Garg And K. Pradhan, "Science Exam Question Answering Based On Retrieval-Augmented Generation," In Proceedings Of The Ieee International Conference On Artificial Intelligence And Applications, 2024.

[10] A. Chowdhury, R. Banerjee, And S. Gupta, "Leveraging Neural Networks In Retrieval-Augmented Qa," Ieee Transactions On Knowledge And Data Engineering, Vol. 35, No. 5, Pp. 123–135, 2023.