# Agentic RAG

## **Abstract**

This project aims to develop an AI-based question-answering system that integrates PDF content and web search results. Utilizing the LangChain framework, HuggingFaceEmbeddings for embeddings, and a custom retrieval chain, the system is designed to provide accurate and concise responses to user queries.

## **Introduction**

The advent of artificial intelligence has transformed how we interact with information. This project explores the capabilities of AI in providing quick and accurate answers to questions derived from various sources, specifically PDFs and web data. The goal is to streamline the process of information retrieval, making it more efficient for users. LangChain is an innovative framework designed to facilitate the development of applications using language models. It provides tools to connect various data sources and integrates different AI models, making it easier to build complex applications like question-answering systems. By leveraging LangChain, developers can efficiently manage the flow of information between components, enhancing the overall performance of AI applications.

## **Literature Review**

Recent advancements in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP), have significantly transformed question-answering systems. The evolution of sophisticated models capable of understanding and generating human language has led to applications that deliver precise and contextually relevant responses.

A pivotal innovation in this field is Retrieval-Augmented Generation (RAG), which combines retrieval systems with generative models. This hybrid approach enhances accuracy by allowing AI to access external information during response generation, overcoming limitations faced by traditional generative models reliant on their training data.

The concept of Agentic RAG advances this further by enabling systems to autonomously retrieve and process information. Such features allow for dynamic interpretation of user queries, fostering more interactive dialogues. This represents a shift towards AI assistants that are active participants in information retrieval.

## **Methodology**

The project employs the LangChain framework to handle PDF content and web search results. Key steps in the methodology include:

1. **PDF Document Loading**: Using the PyPDFLoader to extract content from uploaded PDF files.
2. **Text Splitting**: Implementing RecursiveCharacterTextSplitter for efficient text handling.

3.**Embedding Generation:** Utilizing HuggingFaceEmbeddings to generate embeddings for the text chunks.

4.**Vector Store Creation**: Creating an InMemoryVectorStore for efficient retrieval of relevant information.

5.**Question-Answering Chain:** Setting up a question-answer chain that combines retrieved contexts for generating responses.

## Results and Analysis

Upon testing, the system demonstrates an ability to accurately retrieve information from both PDF and web sources. The combination of local and web search data enhances the overall accuracy and reliability of responses, showcasing the effectiveness of the integrated approach.

## Conclusion and Future Work

The project successfully establishes a framework for an AI-based question-answering system that combines information from multiple sources. Future work will focus on refining the model, enhancing the retrieval mechanisms, and exploring the integration of additional data sources to improve response accuracy and user experience. Specifically, efforts will be made to enable the system to process images and extract information from multiple PDFs simultaneously, thereby broadening its applicability and enhancing its functionality.

**References**

1. LangChain Documentation
2. HuggingFaceEmbeddings Documentation
3. PyPDF Documentation
4. Groq Documentation
5. SerpApi Documentation
6. attention is all you need (<https://arxiv.org/pdf/1706.03762>)