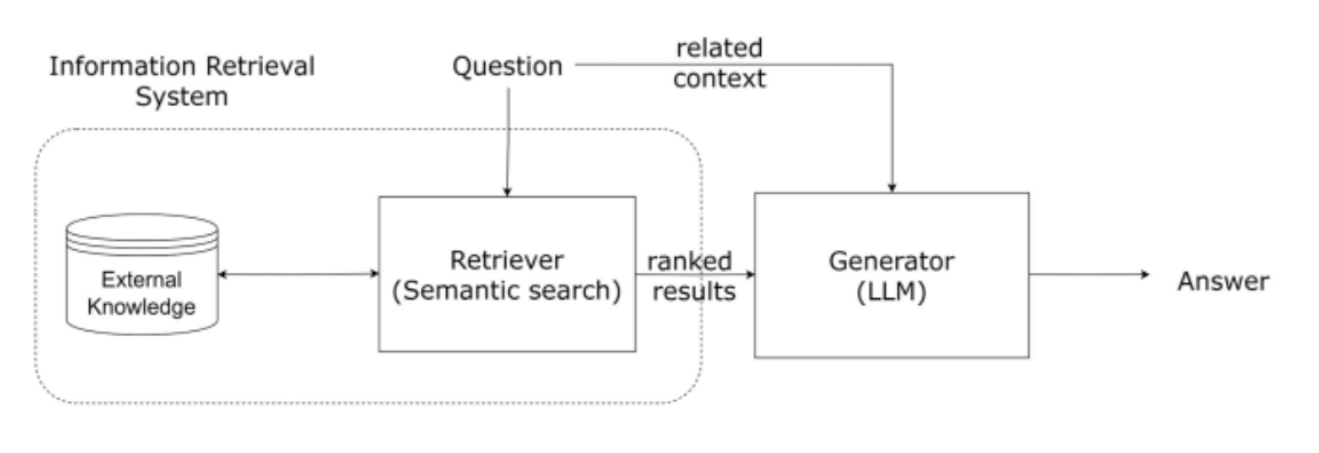
**Semantic Spotter - RAG Project**

**Goal:** Build a robust generative search system capable of effectively and accurately answering questions from various policy documents. You may use LangChain or LlamaIndex to build the generative search application.

**Data Source**: Policy documents present in PDF form as below:

https://drive.google.com/drive/folders/1rfztjM0PHMOHcEantmj9\_kYZlJLtNCw3?usp=drive\_link

**System Design:**

****

RAG (Retrieval-Augmented Generation) is a framework in GenAI that combines the strengths of both retrieval-based and generative models to create more efficient AI assistants. It uses a retrieval mechanism to find relevant information from a large dataset and then generates a response based on that information. This approach improves the accuracy and relevance of responses compared to purely generative models.

LangChain and Llama index can be instrumental in building an efficient AI assistant using RAG. LangChain is a blockchain-based language platform that provides secure and scalable language services. It can be integrated with RAG to enhance its language processing capabilities, allowing the AI assistant to understand and generate more complex and nuanced responses.

On the other hand, Llama index is a large-scale index for efficient language model retrieval. It can be used to quickly retrieve relevant information from a large dataset, which is crucial for the retrieval component of RAG. By leveraging the speed and efficiency of Llama index, the AI assistant can provide faster and more accurate responses to user queries. We would be using **Llama Index** for implementing RAG, for efficient searching in insurance policy documents.

Integrating LangChain and Llama index with RAG can significantly improve the performance and efficiency of AI assistants, making them more capable of understanding and responding to user queries in a more natural and human-like manner.

**Further Improvements**:

This model can be further improved by using better/cleaner dataset or utilizing better data pre-processing techniques.

We may also use custom node and LLMs for better results.