Controlled LLM Inference with Embedding-Based Semantic Search

This project implements an intelligent document-based question answering system by integrating controlled Large Language Model (LLM) inference with embedding-based semantic search. The system allows users to upload TXT or PDF documents and dynamically extract textual content for further processing. After extraction, the text is cleaned and divided into smaller chunks to improve handling of large documents and enhance retrieval accuracy.

Each chunk is converted into vector embeddings using a sentence-transformer model, which captures the semantic meaning of the text. These embeddings are stored in a FAISS vector database to enable fast and efficient similarity search. When a user submits a query, the system retrieves the most relevant text segments by comparing embedding similarity.

The retrieved context is provided to a Large Language Model to generate grounded and meaningful responses. Users can control generation parameters such as Temperature, Top-P, and Maximum Tokens to observe different output behaviors. Deterministic inference is demonstrated by disabling probabilistic sampling, ensuring consistent responses for identical inputs.

This project follows a Retrieval-Augmented Generation (RAG) approach, improving response accuracy while minimizing hallucinations. It showcases practical applications of LLM control and semantic search in modern AI-driven information retrieval systems.