Sanskrit Retrieval-Augmented Generation (RAG) System using CPU-Based LLM

Introduction

Sanskrit is one of the oldest classical languages and contains vast philosophical and literary knowledge.

This project implements a Retrieval-Augmented Generation (RAG) system to answer questions from Sanskrit texts using a CPU-based local model.

Problem Statement

Training large models for Sanskrit is difficult due to limited data and resources.

This project solves the problem by retrieving relevant Sanskrit content and generating grounded answers without training a model.

Objectives

- Ingest Sanskrit documents

- Index them efficiently

- Support Sanskrit, transliteration, and English queries

- Retrieve relevant context

- Generate accurate answers using a local LLM

Dataset Description

The dataset contains Sanskrit prose stories and moral tales stored in DOCX format.

Methodology

Documents are ingested, chunked, embedded using multilingual models, indexed using FAISS,

and queried using a CPU-based LLM with retrieved context.

System Architecture

User Query -> Embeddings -> FAISS -> Retrieved Context -> LLM -> Answer -> Streamlit UI

Results

The system retrieves correct Sanskrit passages and generates meaningful explanations efficiently on CPU.

Advantages

- No training required
- Reduced hallucination
- Fully local execution
- Suitable for low-resource languages

Limitations

- Depends on dataset quality
- Slower inference on low-RAM systems

Future Scope

- Sandhi splitting
- Hybrid retrieval
- Chat memory
- UI enhancements

Conclusion

RAG is an effective approach for Sanskrit NLP, enabling accurate and efficient question answering without heavy computation.