Sanskrit Document Retrieval-Augmented Generation (RAG) System

AI/ML Intern Assignment – Technical Report

1. Introduction

The Sanskrit RAG System is an AI-powered document retrieval and generation system designed to work entirely on CPU. It retrieves relevant Sanskrit documents using vector similarity search and optionally generates contextual answers using a lightweight language model.

2. Problem Statement

Accessing relevant information from large Sanskrit text corpora is challenging due to language complexity and lack of modern tooling. The objective is to design a Retrieval-Augmented Generation system that can answer Sanskrit and English queries efficiently.

3. Dataset Description

The dataset consists of Sanskrit textual documents stored in plain text format. These documents include classical Sanskrit stories, prose, and moral narratives. All documents are encoded in UTF-8 and stored locally for offline processing.

4. System Architecture

The system follows a modular RAG architecture: user query → embedding generation → FAISS vector retrieval → context selection → optional language model generation.

5. Preprocessing Strategy

Documents are split into fixed-size chunks to balance retrieval accuracy and performance. Each chunk is embedded using a sentence transformer model before indexing.

6. Retrieval Mechanism

FAISS is used as the vector database to perform similarity search over embedded document chunks. Top-K most relevant chunks are retrieved for each query.

7. Language Model Generation

An optional CPU-compatible language model is used to generate a natural language response based on the retrieved document context. This ensures minimal hardware requirements.

8. Performance Observations

The system processes queries within sub-second latency for small document collections. Retrieval accuracy is satisfactory for factual and story-based queries.

9. Limitations

The system is optimized for small to medium document collections. LLM generation on CPU may introduce additional latency for complex queries.

10. Future Scope

Future enhancements include better Sanskrit-specific embeddings, multilingual expansion, and improved summarization capabilities.

11. Conclusion

The Sanskrit RAG System demonstrates an effective and lightweight approach to document retrieval and generation using modern NLP techniques under strict CPU constraints.