Machine Learning for RAG Systems

Introduction

This document is a dummy educational PDF designed to be used in a Retrieval-Augmented Generation (RAG) project. It contains foundational machine learning concepts commonly referenced by LLM-powered systems.

What is Machine Learning?

Machine Learning (ML) is a subset of artificial intelligence where systems learn patterns from data instead of being explicitly programmed. ML models improve performance as they see more data.

Core ML Paradigms

Supervised Learning: Learning from labeled data.
Unsupervised Learning: Discovering hidden patterns in unlabeled data.
Reinforcement Learning: Learning through rewards and penalties.

Embeddings

Embeddings are dense vector representations of text, images, or other data. In RAG systems, text embeddings allow semantic search over documents.

Vector Databases

Vector databases store embeddings and enable fast similarity search. Examples include FAISS, Pinecone, Weaviate, and Chroma.

Similarity Search

Similarity search retrieves documents whose embeddings are closest to a query embedding, commonly using cosine similarity or Euclidean distance.

RAG Pipeline Overview

1. Ingest documents
2. Chunk text
3. Generate embeddings
4. Store in vector database

5. Retrieve relevant chunks
6. Generate answer using LLM

Why RAG Matters

RAG systems reduce hallucinations, enable up-to-date knowledge access, and allow domain-specific grounding of large language models.

Conclusion

This PDF serves as a placeholder dataset for experimenting with document loading, chunking, embedding generation, and retrieval in a RAG system.