AI Study Assistant

Backend Development & Architecture

Graduation Project

Team of 6 Students | Date: November 29, 2025

Project Idea

The Problem

Students often struggle to find specific answers within vast amounts of study material, leading to inefficient study sessions and information overload.

Proposed Solution

A "Local-first" Retrieval-Augmented Generation (RAG) system. It ingests PDF documents, creates local embeddings, and uses LLMs to provide precise answers, summaries, and Q&A pairs.

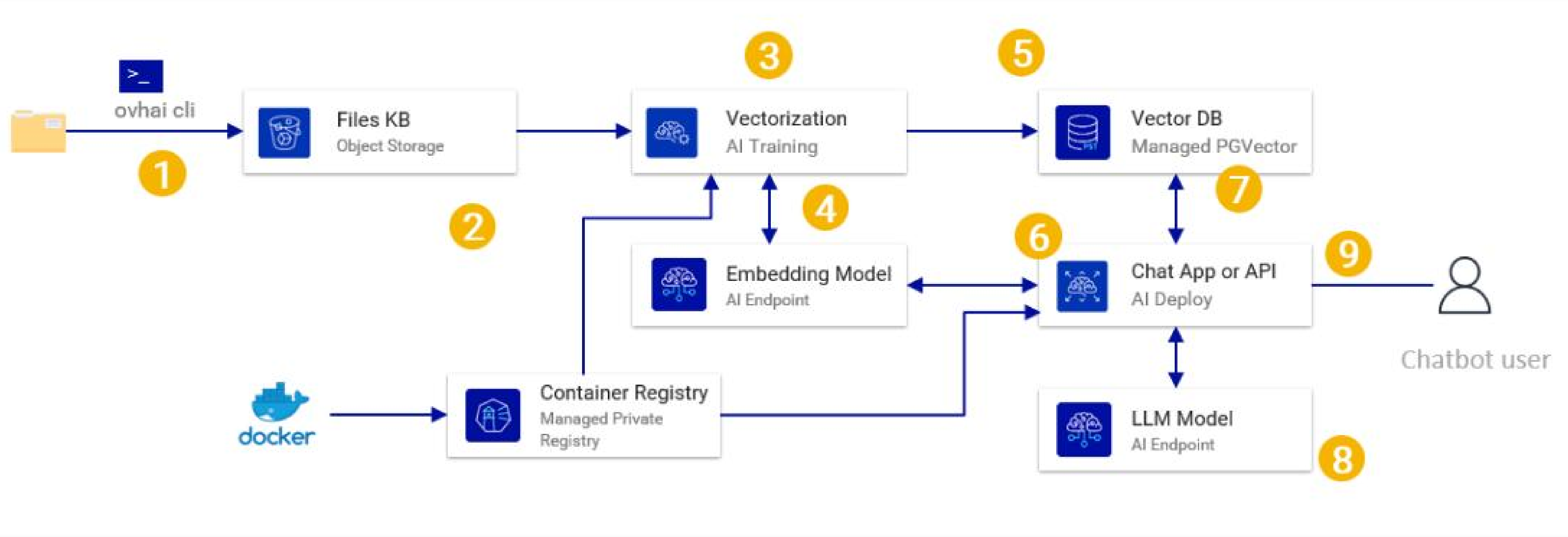
Unique Value Proposition

Unlike cloud-only solutions, our system prioritizes **Data Privacy** and **Offline Capability** by utilizing FAISS and local models (Ollama), falling back to Gemini only when necessary.

System Architecture

Backend Workflow

- ✓ **Input:** User uploads PDF documents via FastAPI endpoints.
- ✓ **Processing:** `pdf_processor.py` extracts text and chunks data.
- ✓ **Retrieval:** `improved_rag_retriever.py` creates vectors using sentence-transformers and stores them in FAISS.
- ✓ **Inference:** The system queries the Local LLM (Ollama) or Cloud Fallback (Gemini) to generate context-aware answers.



End Users & Key Features

Target Audience

Students: Need quick answers from textbooks.

Researchers: Need to summarize long papers.

Educators: Auto-generate Q&A for quizzes.

Core Features

RAG QA: Ask questions directly to your documents.

Summarization: BART-large-CNN hierarchical synthesis.

Q&A Generation: FLAN-T5 automated question creation.

User Benefits

Privacy: Data stays local on the device.

Efficiency: Saves hours of manual reading.

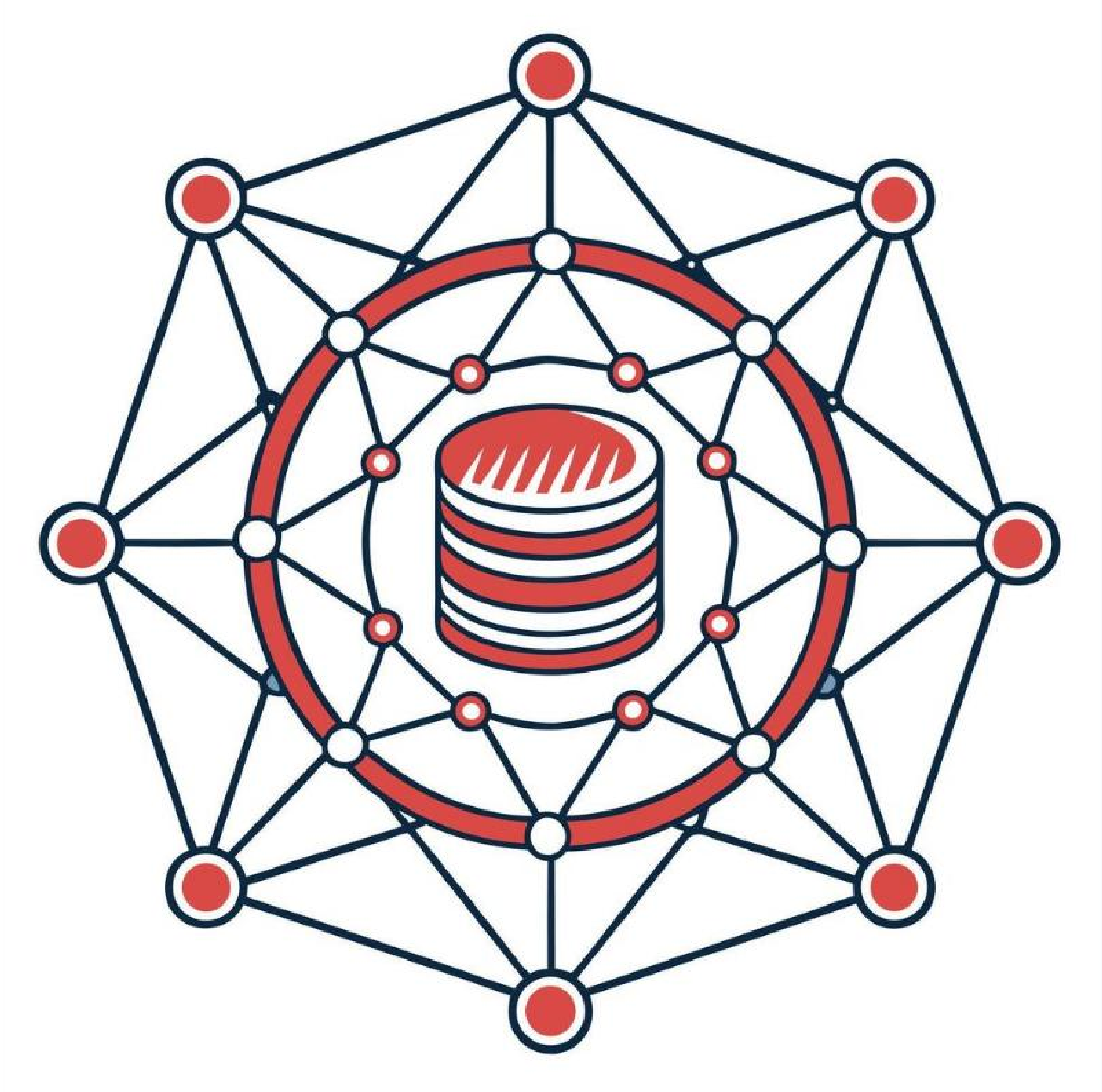
Flexibility: Works with or without internet.

Data Structure & Flow

Vector Database Architecture

We utilize a non-relational, vector-based approach for document retrieval.

- ✓ **Storage:** FAISS (Facebook AI Similarity Search) index.
- ✓ **Entities:** Document Chunks (Text), Vector Embeddings (Float32 arrays), Metadata (Source, Page #).
- ✓ **Data Flow:**
  1. PDF Upload
  2. Text Extraction & Chunking
  3. Embedding Generation (Sentence-Transformers)
  4. Indexing in FAISS
  5. Query & Retrieval



Tech Stack

Python

Core language for backend logic and AI integration.

FastAPI

High-performance web framework for building APIs.

LangChain

Framework for chaining LLM components and retrieval.

PyTorch

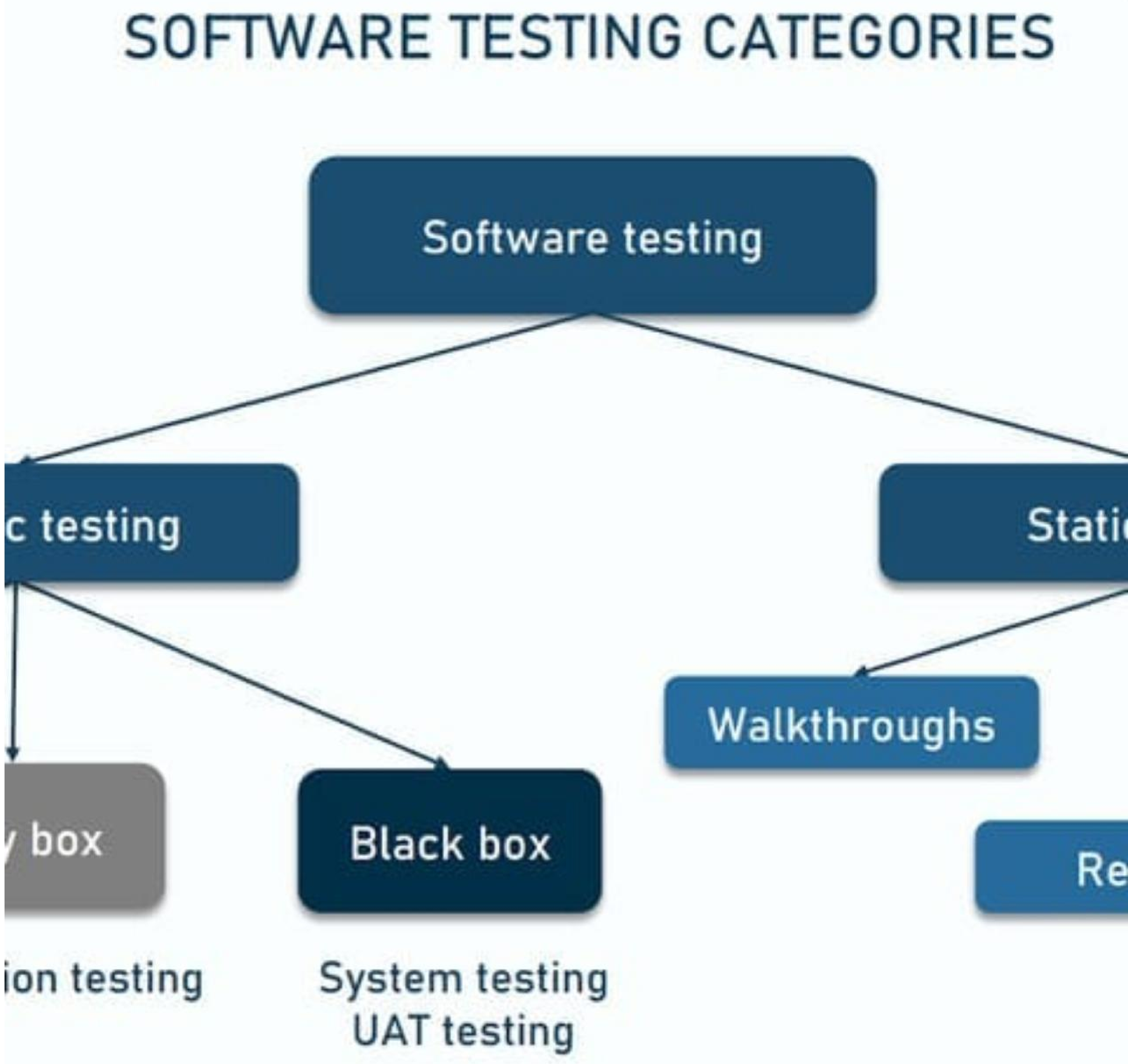
Underlying library for Transformer models.

Testing & Validation

Current Status: Beta Release. Core RAG functionality and API endpoints are fully operational.

Testing Phases

- ✓ **Unit Testing:** PyTest for individual modules (PDF processing, Summarizer).
- ✓ **Integration Testing:** Verifying data flow from Upload -> Vector Store -> LLM Response.
- ✓ **Performance Testing:** Measuring retrieval latency and generation time on local hardware.



Project Deliverables

Technical Documentation

Comprehensive API documentation (Swagger/Redoc), setup guides, and architecture specs.

Source Code

Full GitHub repository with clean, modular code for the Backend system.

User Manual

Guide on how to configure local models (Ollama) and manage API keys.

Final Report

The Team

Member 1

Backend Lead

Member 2

AI Engineer

Member 3

DevOps & Cloud

Member 4

Data Engineer

Member 5

QA Specialist

Member 6

Documentation

Thank You

We are open for your questions.

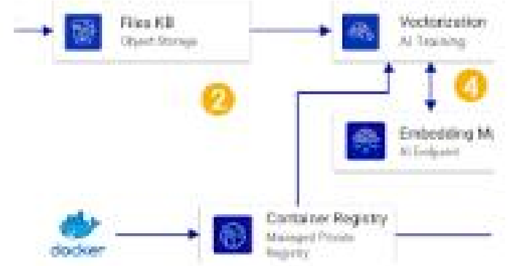
✉ contact@depi-project.edu

Ministry of Communications and Information Technology - DEPI

Image Sources

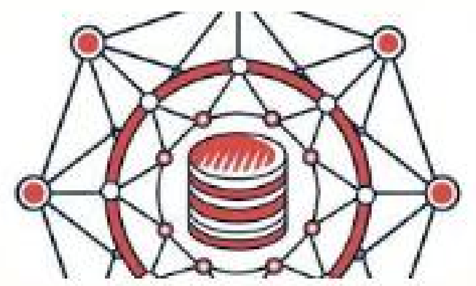
https://news.mit.edu/sites/default/files/styles/news_article__image_gallery/public/images/202011/artificial-intelligence-3382521_1920.jpg?itok=ARma5ST6

Source: news.mit.edu



https://blog.ovhcloud.com/wp-content/uploads/2024/09/80604d005ee408fb2398ed01815737e8-img_34.png

Source: blog.ovhcloud.com



https://static.vecteezy.com/system/resources/previews/070/470/687/non_2x/abstract-network-graph-visualization-with-central-database-icon-representing-data-connectivity-and-information-flow-free-vector.jpg

Source: www.vecteezy.com

https://www.altexsoft.com/static/blog-post/2023/12/025a2634-2677-4fd0-a970-4c93e607a75c.jpg

Source: www.altexsoft.com