YouTube RAG Chatbot - System Architecture

This document describes the system architecture for the YouTube RAG Chatbot. The system is composed of a frontend built with React, a backend using FastAPI, and a set of AI services and tools including LangChain, FAISS, Google Generative AI API, and YouTubeTranscriptAPI.

# Architecture Overview

The following components make up the architecture:

* 1. Frontend (React + Vite): Accepts YouTube URL and user questions, communicates with backend.
* 2. Backend (FastAPI): Processes input, interacts with external APIs, performs RAG-based inference.
* 3. YouTubeTranscriptAPI: Fetches transcript of the video.
* 4. Embedding Generator (Google Generative AI Embeddings): Converts text into vector embeddings.
* 5. Vector Store (FAISS): Stores embeddings and retrieves relevant ones based on user queries.
* 6. LangChain: Constructs RAG pipeline and manages LLM interaction.
* 7. LLM (Gemini 2.5 Flash): Generates natural language answers using retrieved context.
* 8. Communication Layer: RESTful APIs between frontend and backend.

# Data Flow

1. User enters a YouTube URL and a question in the frontend.  
2. Frontend sends this data to the FastAPI backend.  
3. Backend uses YouTubeTranscriptAPI to fetch the transcript.  
4. Transcript is chunked and embedded using Google Generative AI Embeddings.  
5. Embeddings are stored and retrieved using FAISS based on question similarity.  
6. Retrieved chunks are sent to Gemini 2.5 Flash via LangChain.  
7. Gemini generates an answer which is returned to the frontend.

# Architecture Diagram

