React Frontend (UI/UX Layer) Input: YouTube URL Input: User Question Output: Displays Al Response Sends: POST request to FastAPI Displays: Loading state Displays: Error messages FastAPI Backend (Controller Layer) Receives: input from React Validates: URL and question Orchestrates: processing flow Parses YouTube URL to extract video ID Calls YouTube API for transcript Handles errors/ missing transcripts Calls LangChain pipeline Sends Al answer back to frontend LangChain RAG Pipeline YouTube API Splits transcript using TextSplitter Get Video ID Embeds using Google API Calls YouTubeTranscriptAPI Stores vectors in FAISS Returns subtitles · Performs similarity search Handles unavailable or auto-· Sends context to LLM generated captions **Google Generative AI Embedding Model** • Model: embedding-001 · (Accessed via LangChain) **FAISS Vector Store** Stores vectors Returns top-k relevant chunks Gemini 2.5 Flash (LLM) Vector search via cosine similarity Receives question + top context Returns natural language answer Can handle multimodal (future)