Studio Tours AI Assistant - Simplified Architecture

How the System Works (End-to-End)

⏱ **Average Response: < 3 seconds**

User

Speaks/Types Question

AI Processing Engine

1️⃣ **Understand Query**

2️⃣ **Search Knowledge**
Vector + Keyword Search

3️⃣ **Rank Results**
Cross-Encoder Reranking

4️⃣ **Generate Answer**
Context + LLM (Llama 3.1)

Web Interface

🎤 Voice Input

💬 Chat UI

⚙ Settings

Answer

Query

Process

AI Response

Streaming answer with

source citations

+ Suggested questions

⚡ *Real-time*

Key Features

⚡ Streaming responses

🔍 Hybrid search

🔒 100% local & private

🧠 Context-aware

🌐 Multi-language

Search

Knowledge Base

📖 Studio Tour Guides *2000+ chunks*

🎬 Production History *Vector + Keyword*

❓ FAQs & Policies

Next.js   **FastAPI**   **ChromaDB**   **Llama 3.1**

Data flows: User → Frontend → AI Engine ↔ Knowledge Base → Response