1. Project Overview

This project is a real-time voice chatbot that enables natural voice conversations using fully local AI models.

---

2. System Workflow

1. User speaks into the microphone.
2. FastRTC captures and streams the audio.
3. Moonshine model converts speech to text (Speech-to-Text).
4. Ollama runs the Gemma language model to generate a response.
5. Kokoro converts the response text into speech (Text-to-Speech).
6. Audio is streamed back to the user in real time.

All processing happens locally.

---

3. Core Technologies

- FastRTC – Real-time WebRTC communication
- Moonshine – Speech-to-Text model
- Ollama – Local LLM runtime
- Gemma (1B/4B) – Language model
- Kokoro – Text-to-Speech model
- Gradio – Web-based user interface

4. Data Flow Summary

Voice Input → Audio Capture → Speech-to-Text → LLM Processing → Text-to-Speech → Audio Output