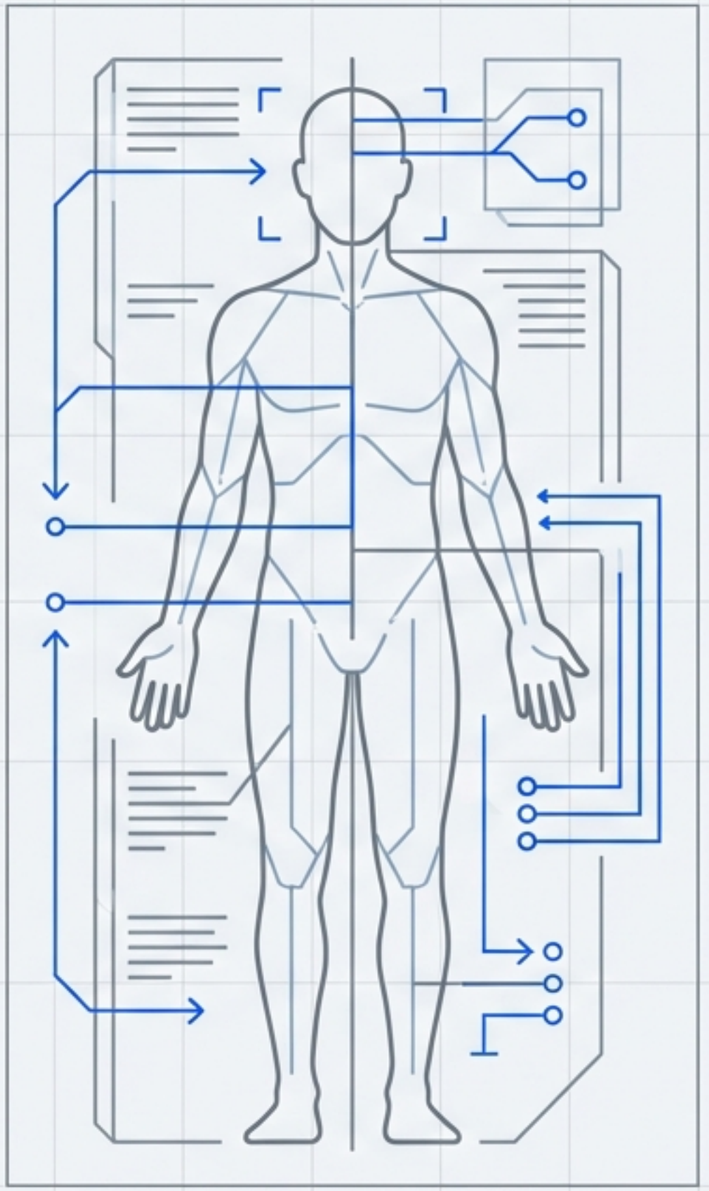
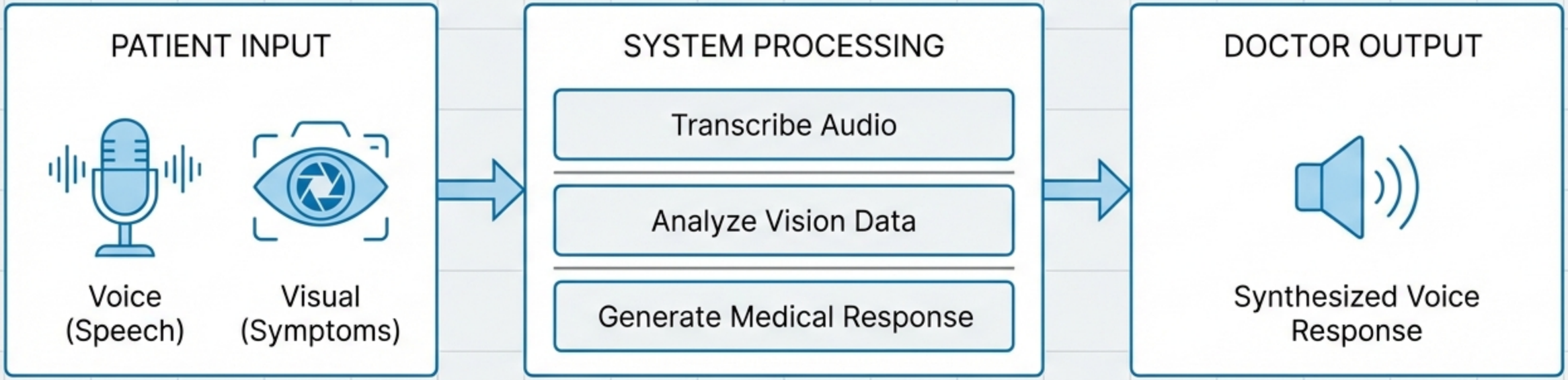
AI DOCTOR 2.0

Building a Multimodal Medical Chatbot with Vision and Voice.



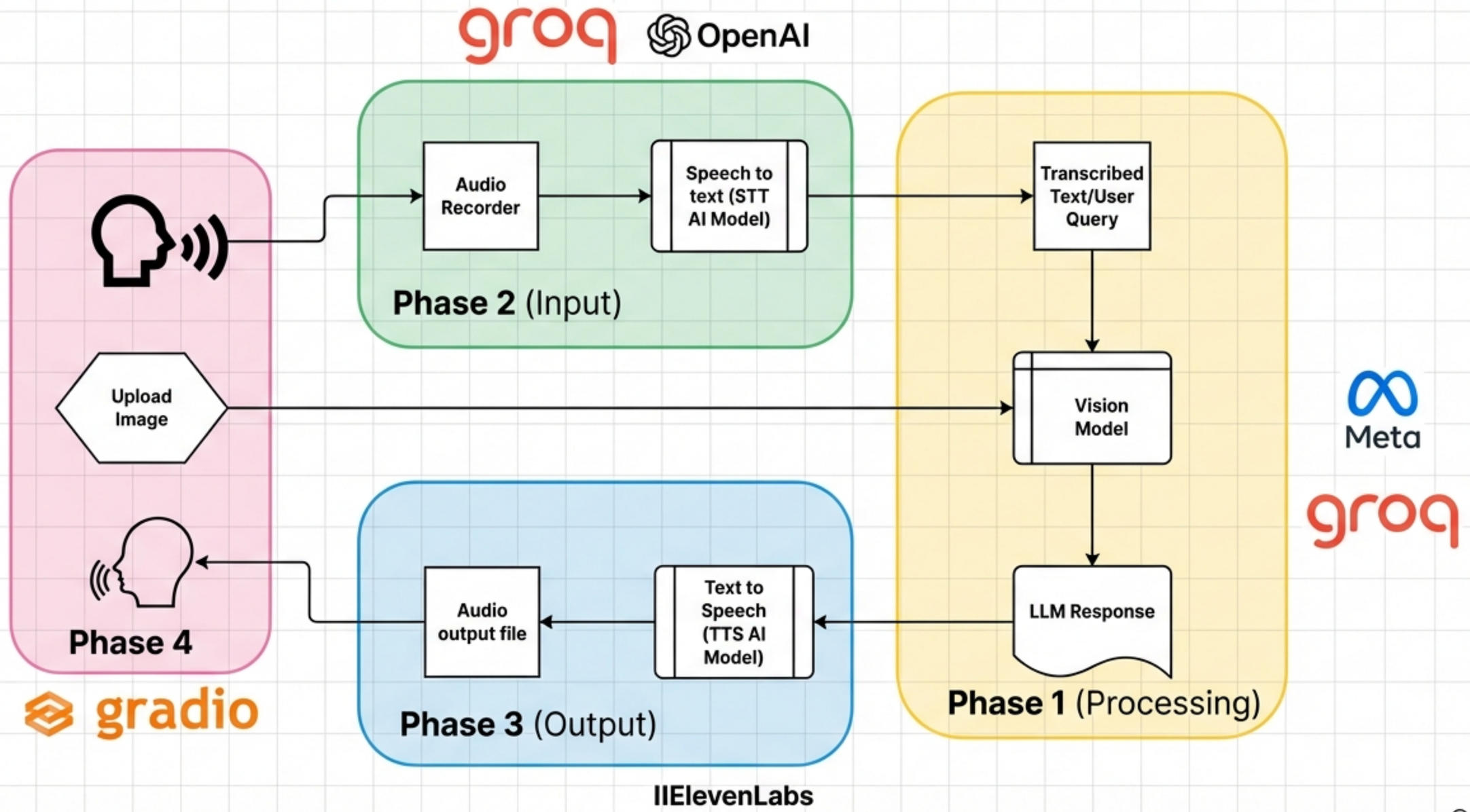
SYSTEM ARCHITECTURE: MULTIMODAL LLM INTEGRATION // GROQ + LLAMA 3 + WHISPER

THE VISION: A DOCTOR THAT SEES AND HEARS



A multimodal consultation loop mimicking real-world interaction: analyzing physical symptoms and verbal descriptions simultaneously.

TECHNICAL ARCHITECTURE: THE BLUEPRINT



THE TECH STACK

Inference Engine

Groq
JetBrains Mono
Low-latency AI inference
Inter

The Brain

Llama 3 Vision
JetBrains Mono
Open source Multimodal LLM
Inter

The Ears

OpenAI Whisper
JetBrains Mono
High-accuracy transcription
Inter

The Voice

gTTS & ElevenLabs
JetBrains Mono
Text-to-Speech synthesis
Inter

The Interface

Gradio
JetBrains Mono
Rapid UI deployment
Inter

Environment

Python / VS Code
JetBrains Mono
Core development platform
Inter

FOUNDATION: AUDIO DEPENDENCIES

Prerequisite setup for FFfmpeg and PortAudio

macOS

1. Install Homebrew
2. Command:

```
brew install ffmpeg portaudio
```

Linux (Debian/Ubuntu)

1. Update package list
2. Command:

```
sudo apt-get install ffmpeg libportaudio2
```

Windows

1. FFmpeg: Download static build, extract to C:\ffmpeg, add /bin to PATH.
2. PortAudio: Download and install official binaries.

FOUNDATION: PYTHON ENVIRONMENT

Pipenv

```
pipenv install →
pipenv shell
```

Venv

```
python -m venv venv →
pip install -r requirements.txt
```

Conda

```
conda create --name ai-doctor
```

Critical Dependencies

```
groq==0.15.0
gradio==5.12.0
speechrecognition==3.13.0
elevenlabs==1.50.3
pydub==0.25.1
```

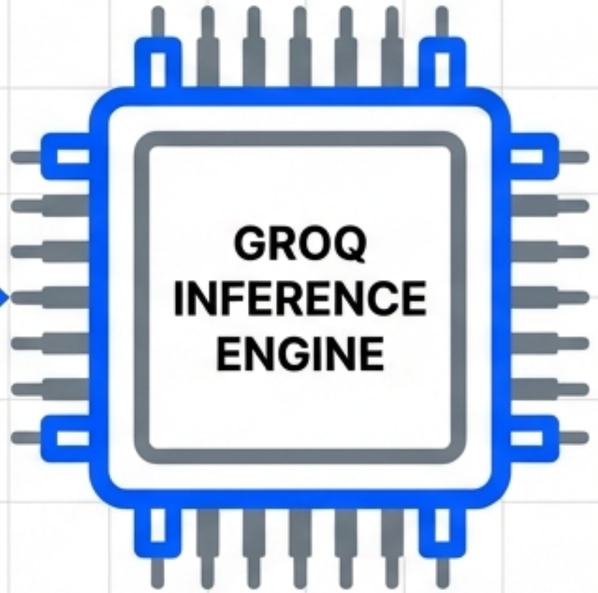
PHASE 1: THE BRAIN (MULTIMODAL LLM)

Configuration

- Set `GROQ_API_KEY`
- Initialize Llama 3 Vision Model

Data Formatting

Base64 / Tensor Conversion



The core intelligence loop. Images are encoded and sent to Groq's Llama 3 endpoint alongside the user's text query to generate a diagnosis.

PHASE 1 IN ACTION: VISION CAPABILITIES

MEDICAL RECORD PHOTO

DIAGNOSTIC REPORT

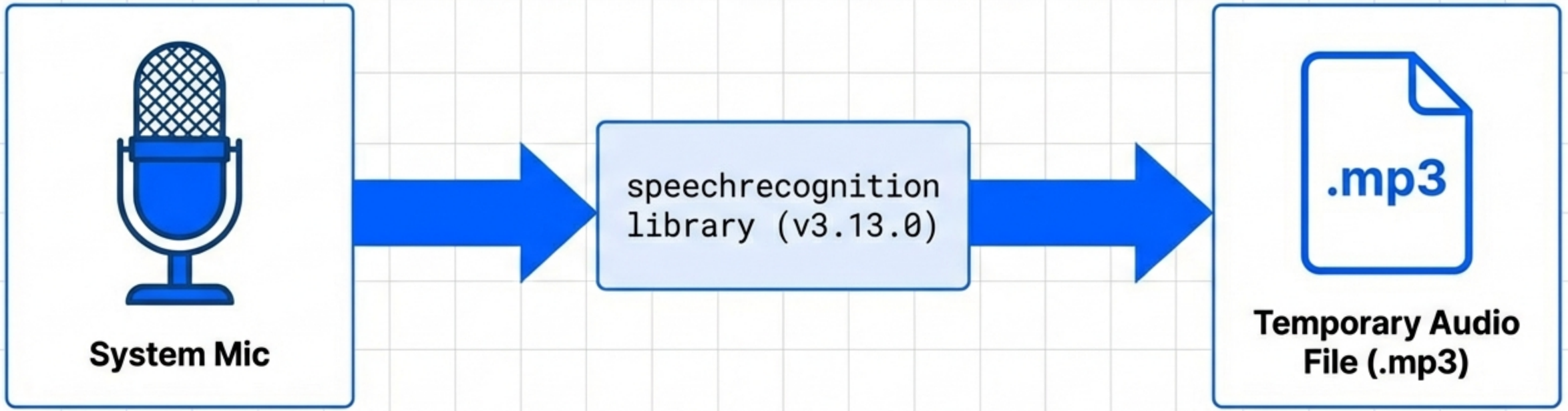
INPUT QUERY

Analyze this image and identify potential conditions.

LLAMA 3 VISION OUTPUT

Observation: Redness and irritation visible on the forearm. Texture appears consistent with contact dermatitis or mild eczema.
Recommendation: Monitor for spreading; consult a dermatologist if itching persists.

PHASE 2: VOICE OF THE PATIENT



Dependencies

PortAudio & PyAudio used to interface with hardware.

Side Note

Verification: Validated via `patient_voice_test.mp3`.

Swiss Clinical Architectural

PHASE 2 DETAIL: TRANSCRIPTION (STT)

Input Audio

OpenAI Whisper Model

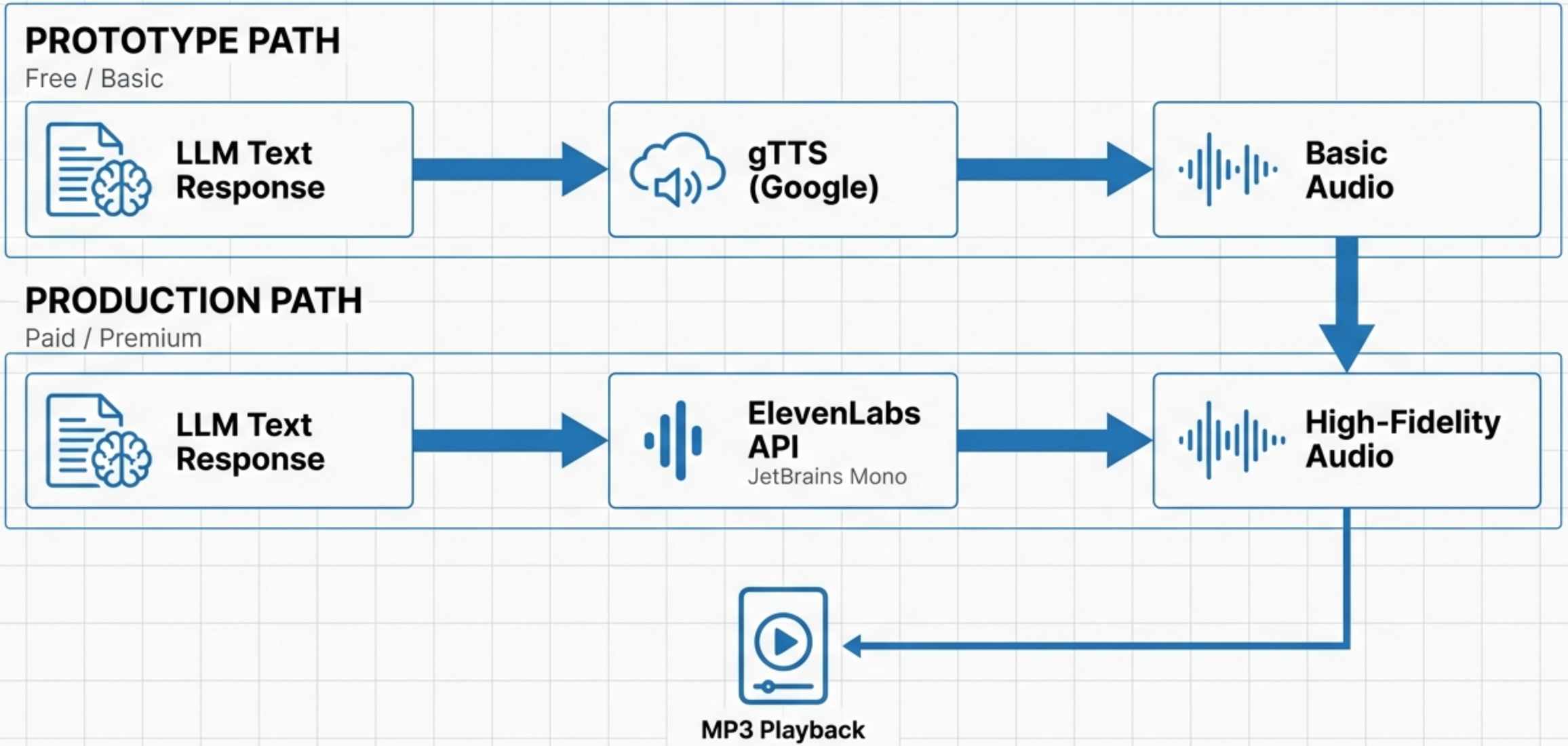
`Whisper Model`

Output Text

The Transcribed Text serves as the prompt for the Phase 1 Vision Model.

Selected as the best open-source model for transcription.

PHASE 3: VOICE OF THE DOCTOR (TTS)



PHASE 3 ANALYSIS: MODEL COMPARISON

gTTS

Type: Open Source / Free

Quality: Robotic, flat intonation

Use Case: Testing & Debugging

File Ref: `gtts_testing.mp3`

ElevenLabs

Type: Commercial API

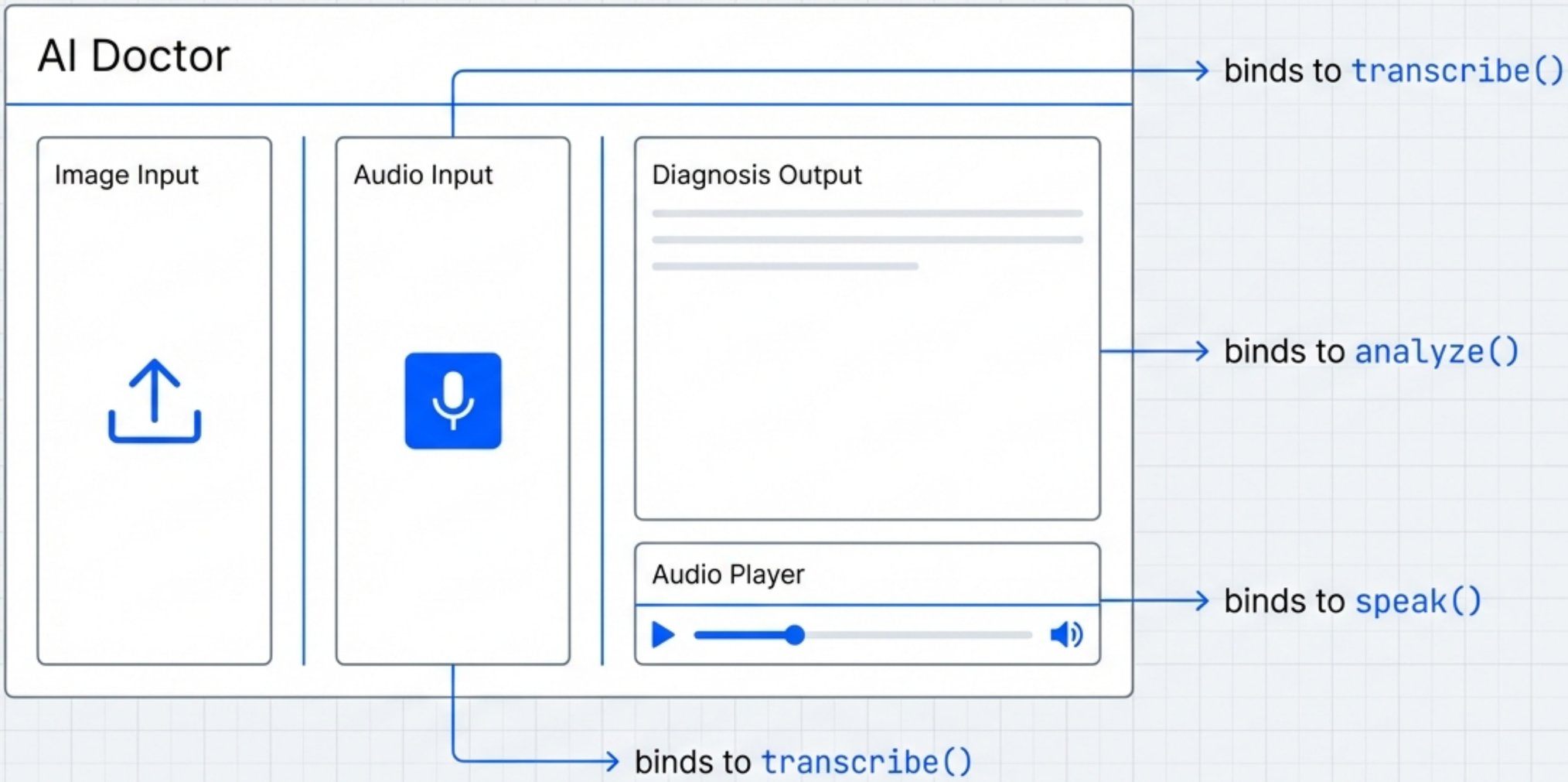
Quality: Natural, authoritative, human-like

Use Case: Final User Experience

File Ref: `elevenlabs_testing.mp3`

Start with `gTTS` for the build, swap to `ElevenLabs` for the demo.

PHASE 4: THE INTERFACE (GRADIO UI)



FUTURE ROADMAP & IMPROVEMENTS

Medical Accuracy

Fine-tune vision models on dermatology datasets.

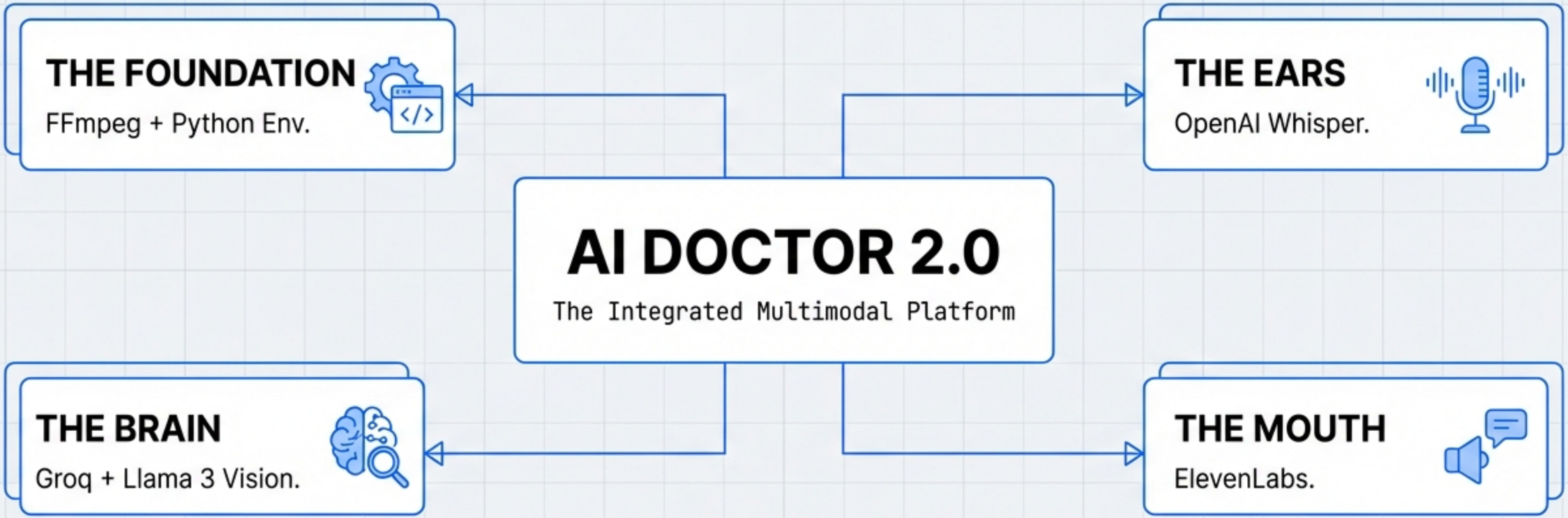
Model Performance

Upgrade to state-of-the-art Paid LLMs for higher precision.

Accessibility

Implement multilingual support for global patient reach.

SUMMARY: THE COMPLETE SYSTEM



READY TO DEPLOY

A fully functional multimodal assistant bridging vision and voice.