AI DOCTOR

(Vision and Voice)

Medical Chatbot with MultiModal LLM

NAME – ABHISHEK YADAV

MASTER'S COMPUTER SCIENCE(AI)

DEEP LERANING FOR NPL
PROJECT SUMMARY

PROJECT OVERVIEW

In this capstone project, I have created an Industry-Specific Large Language Model (LLM) Bot using state-of-the-art pre-trained models from sources like GROQ, LLaMA4 and ELEVENLABS. The primary objective is to build an intelligent bot that can effectively engage with users by answering questions and providing insights specific to a chosen industry. This project will not only enhance your technical skills but also provide a deep understanding of the chosen industry's nuances, challenges, and trends.

This project creates a voice-interactive AI doctor that:

•Takes **voice input** from the user (the patient).
•Converts speech to text using **Groq's Whisper** model.
•Combines this text with an optional uploaded **medical image**.
•Analyzes the image and query using a **Vision-Language model**.
•Responds with a **diagnosis or advice** as a **realistic doctor**.
•Converts this response into **spoken audio** using ElevenLabs' Text-to-Speech.

PROJECT LAYOUT

Phase 1–Setup the brain of the Doctor (Multimodal LLM)

- Setup GROQ API key
- Convert image to required format
- Setup Multimodal LLM

Phase 2–Setup voice of the patient

- Setup Audio recorder (ffmpeg & portaudio)
- Setup Speech to text–STT–model for transcription

Phase 3–Setup voice of the Doctor

- Setup Text to Speech–TTS–model (ElevenLabs)
- Use Model for Text output to Voice

Phase 4–Setup UI for the VoiceBot

- VoiceBot UI with Gradio

Phase 1–Setup the brain of the Doctor (Multimodal LLM)

Purpose:
Processes medical images and generates intelligent analysis using a Vision-Language model.

Key Components:

- **Encode_image(image_path):**
  Loads the image and encodes it into base634 format.
- **Analyze_image_with_query(query, encoded_image, model):**
  Sends query and image to a model hosted on Hugging Face inference API
  e.g., (meta-llama/llama-4-scout-17b-16e-instruct).
- **Return a response in natural language.**

```
# Step 3: Setup Multimodal LLM
query = "Is there something wrong with my face?"
model = "meta-llama/llama-4-scout-17b-16e-instruct"

# Passing the key from env

def analyze_image_with_query(query, model, encoded_image):
    client=Groq()
    messages=[
```

Phase 2–Setup voice of the patient

Purpose:
Captures audio input and transcribes it into text using Groq's Whisper model.

Key Components:
- record_audio(): Allows microphone input recording (used in UI).
- transcribe_with_groq():
  Uploads the audio to Groq API.
  Uses whispers-large-v3 to convert speect to text.

```
GROQ_API_KEY = os.getenv("GROQ_API_KEY")
stt_model="whisper-large-v3"

def transcribe_with_groq(stt_model, audio_filepath, GROQ_API_KEY):
    client=Groq(api_key=GROQ_API_KEY)
```

Phase 3–Setup voice of the Doctor

Purpose:
Converts text response (diagnosis or advice) into a voice response using ElevenLabs.

Key Component:
Text_to_speech_with_elevenlabs(input_text, output_filepath):
Sends input_text to Elevenlabs using API key and voice ID.
Saves spoken output as MP3.

```
input_text="Hi this is Ai medical bot, autoplay testing!"

def text_to_speech_with_elevenlabs(input_text, output_filepath):
    client=ElevenLabs(api_key=ELEVENLABS_API_KEY)
    audio=client.generate(
        text= input_text,
        voice= "Brian",
        output_format= "mp3_22050_32",
        model= "eleven_turbo_v2"
```

Phase 4–Setup UI for the VoiceBot

Purpose:

Integrates all components into a single user-friendly interface.

Key Component:

Takes voice and image as input

Calls:

transcribe_with_groq()

analyze_image_with_query()

text_to-speech_with_elevenlabs()

Shows output text and plays voice

```
def process_inputs(audio_filepath, image_filepath):
    speech_to_text_output = transcribe_with_groq(GROQ_API_KEY=os.getenv("GROQ_API_KEY"),
                                                 audio_filepath=audio_filepath,
                                                 stt_model="whisper-large-v3")
```

```
from brain_of_the_doc import encode_image, analyze_image_with_query
from voice_of_patient import record_audio, transcribe_with_groq
from voice_of_doctor import text_to_speech_with_elevenlabs
```

127.0.0.1:7860

audio_filepath

0:07 0:07

1x

image_filepath

Speech to Text

can you explain what is the cause of these dark spots on my face

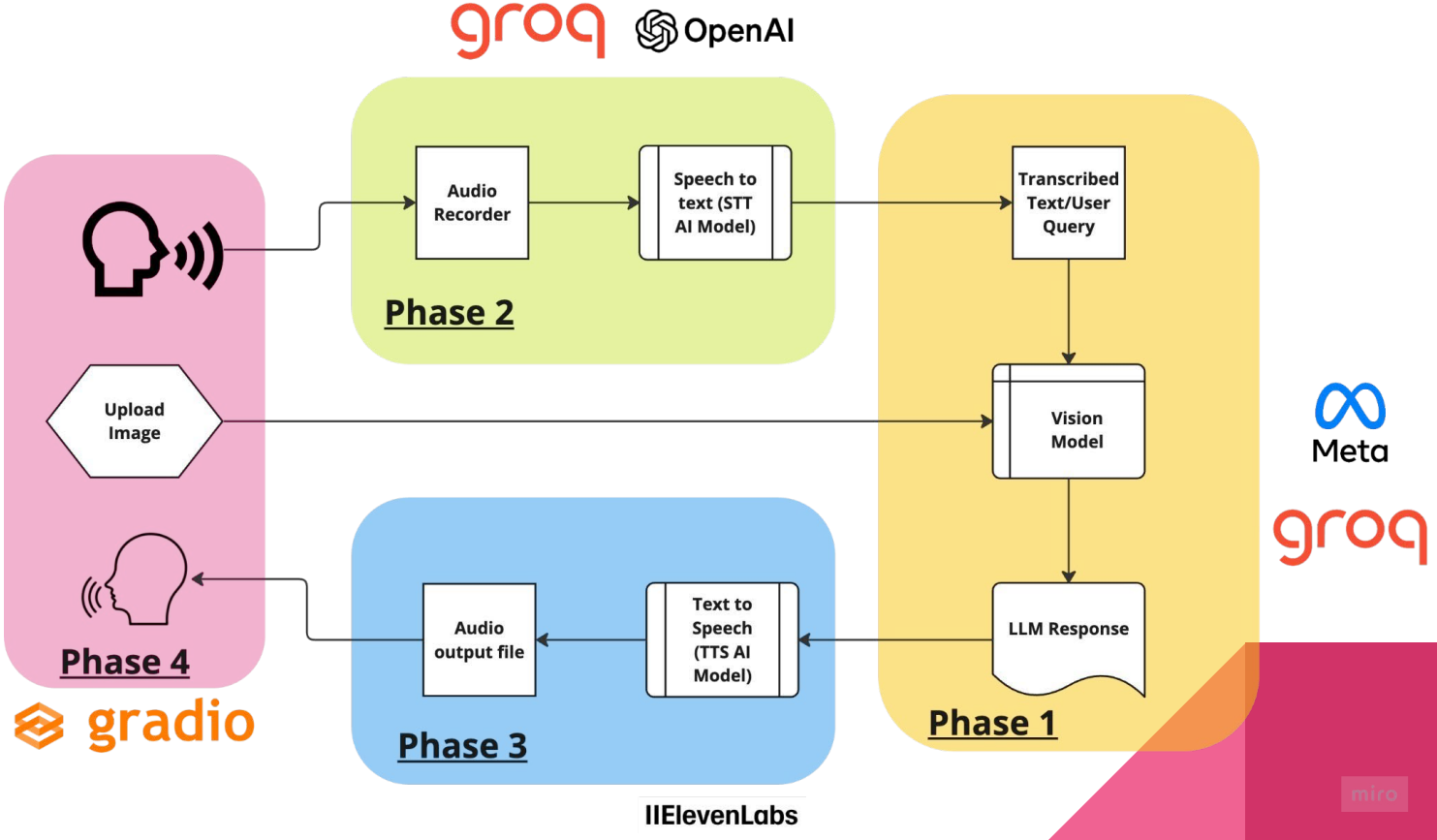
Doctor's Response

With what I see, I think you have hyperpigmentation, which is a common skin condition characterized by dark spots or discoloration on the skin. This can be caused by a variety of factors, including sun exposure, genetics, hormonal changes, and certain medications. To address this issue, I would recommend using topical creams or serums containing ingredients such as hydroquinone, kojic acid, or niacinamide, which have been shown to be effective in reducing hyperpigmentation. Additionally, protecting your skin from further sun damage by using sunscreen with at least SPF 30 and wearing protective clothing can help prevent further darkening of the spots.

output 2

Flag

TECHNICAL ARCHITECTURE



TOOLS AND TECHNOLOGIES

- Groq for AI Inference
- OpenAI Whisper (Best open source model for Transcription)
- Llama 3 Vision (Open source by Meta)
- ElevenLabs (Speech to Text)
- Gradio for UI
- Python
- VS Code