Dr. Freddy: The Virtual AI Doctor

An AI-Powered Multimodal Conversational Agent for Preliminary Medical Interaction

Project Report

By

Atharv Ashok Mujumale

An Independent Project Submitted in Support of Application for Master's Programs

October 2025

Abstract:

This project introduces "Dr. Freddy: The Virtual AI Doctor," a novel multimodal AI conversational agent designed to facilitate preliminary medical interactions. Leveraging a robust architecture, Dr. Freddy integrates voice and image input from users (patients) to understand their health concerns. The system utilizes OpenAI Whisper for accurate speech-to-text transcription, a Meta Llama-based large language model (LLM) hosted on Groq for intelligent medical reasoning and response generation, and ElevenLabs for natural-sounding text-to-speech conversion. The user interface, built with Gradio, offers an intuitive and interactive experience, enabling users to record their voice, upload relevant images, and receive both textual and auditory responses. Engineered with a custom doctor persona, Dr. Freddy aims to provide accessible and empathetic preliminary medical guidance, showcasing the transformative potential of AI in healthcare. This report details the system's design, implementation, and evaluates its current capabilities as a foundation for future advancements in AI-driven medical consultation.

Chapter 1: Introduction

1.1. Motivation & Background

The rapid advancements in Artificial Intelligence (AI), particularly in natural language processing and computer vision, have opened unprecedented opportunities across various sectors. One of the most promising yet challenging fields for AI application is healthcare. Traditional healthcare systems often face challenges related to accessibility, time constraints, and the need for immediate preliminary advice, especially in remote areas or during off-hours. Virtual assistants and conversational AI agents have emerged as potential solutions to bridge these gaps, offering preliminary consultations, answering frequently asked questions, and guiding users to appropriate care. However, many existing solutions are primarily text-based or lack the ability to process diverse forms of input, such as images, which are often crucial for medical diagnosis and understanding symptoms. This project is motivated by the vision of creating a more comprehensive, interactive, and human-like AI interface for initial medical interactions, thereby enhancing patient engagement and improving the efficiency of preliminary healthcare assessments.

1.2. Problem Statement

Current AI solutions for health-related queries often fall short in providing a holistic interactive experience. They typically rely solely on text input or lack the ability to incorporate visual information, which is critical for describing symptoms like rashes, injuries, or other visible conditions. Furthermore, the absence of natural voice interaction can make the experience feel impersonal and less intuitive for users, especially those who prefer verbal communication. The core problem addressed by this project is the lack of an easily accessible, multimodal (voice + image), and voice-responsive AI agent capable of acting as a preliminary virtual doctor, understanding complex patient inputs, and generating empathetic, medically-oriented responses.

1.3. Project Objectives

The primary objectives of "Dr. Freddy: The Virtual AI Doctor" are:

- To design and implement a multimodal AI conversational agent capable of accepting both voice and image inputs from users.
- To accurately transcribe spoken user queries into text using advanced speech-to-text models.
- To develop a robust AI reasoning core that processes both textual and visual information to understand medical scenarios.
- To generate coherent, medically-relevant, and empathetic responses, maintaining a professional doctor persona.
- To convert generated text responses into natural-sounding speech for an enhanced user experience.
- To create an intuitive and interactive graphical user interface (GUI) for seamless user interaction, including recording audio, uploading images, and displaying responses.
- To demonstrate the feasibility and potential of integrating various cutting-edge AI services (LLMs, STT, TTS, Vision) into a unified application for healthcare preliminary consultation.

1.4. Scope and Limitations

The scope of this project is to develop a proof-of-concept, working model of a multimodal virtual AI doctor. Dr. Freddy is designed to provide *preliminary* medical interaction and guidance, acting as an initial point of contact for users to describe their symptoms and receive general information. It is explicitly not intended to replace professional medical diagnosis, treatment, or advice from qualified human healthcare providers. The current model's accuracy is dependent on the underlying AI models and their training data, and while a doctor persona is enforced, the depth of medical knowledge is constrained by the general-purpose nature of the LLM used (Meta-Llama). Future improvements could involve fine-tuning the LLM on specific medical datasets to enhance diagnostic accuracy and personalized advice. The project utilizes free and open-source alternatives, which may have performance or feature limitations compared to dedicated commercial medical AI platforms.

Chapter 2: Literature Review & Technology Stack

2.1. Literature Review

The conceptual foundation for "Dr. Freddy" is informed by the growing body of research on the application of Artificial Intelligence in healthcare. A pivotal article, "Artificial Intelligence in Medicine" published in the National Center for Biotechnology Information (NCBI), highlights the transformative potential of AI in diagnostics, treatment, and patient interaction. The paper discusses how AI algorithms can analyze complex medical data, from medical images to patient records, to assist clinicians in making more accurate and timely decisions. This concept of AI as a supportive tool for medical professionals is a core principle behind Dr. Freddy, which aims to act as an initial point of contact, thereby streamlining the diagnostic process.

Furthermore, the project builds upon the trends observed in the broader software development community, where open-source technologies and accessible AI models have democratized the creation of sophisticated applications. Practical implementation and debugging techniques were refined by consulting various educational resources, including programming tutorials from online creators like "CodeWithHarry" on YouTube. These resources provided valuable insights into the practical aspects of integrating different APIs and libraries, which is a common challenge in building multimodal AI systems. Dr. Freddy thus sits at the intersection of academic inspiration and hands-on, community-driven technical execution.

2.2. Technology Stack

The selection of technologies for this project was guided by the principles of using powerful, open-source or freely accessible tools that allow for rapid prototyping and robust performance. Each component was chosen for its specific strengths in the overall architecture.

- **Programming Language: Python**
    - **Justification:** Python was chosen as the primary development language due to its extensive ecosystem of libraries for AI, machine learning, and web development. Its simple syntax and strong community support make it ideal for rapid application development.
- **Development Environment: Visual Studio Code**
    - **Justification:** VS Code provided a versatile and lightweight environment with excellent support for Python, debugging tools, and terminal integration, which streamlined the development workflow.
- **AI Inference & Model Hosting: Groq**
    - **Justification:** Groq was selected for its high-performance inference engine, offering fast and efficient access to powerful, open-source Large Language Models. Its API-first approach simplified the integration of complex models into the application.
- **Core AI Brain: Meta-Llama-4-scout-17b-16e-instruct**
    - **Justification:** This state-of-the-art multimodal language model from Meta was chosen for its ability to process both text and images simultaneously. This was a critical requirement for fulfilling the project's core objective of analyzing patient-provided images alongside their verbal descriptions.
- **Speech-to-Text (STT): OpenAI Whisper**
    - **Justification:** Accessed via the Groq API, Whisper (large-v3 model) is renowned for its high accuracy in transcribing spoken language, even with various accents and background noise. This ensures that the patient's spoken query is captured precisely for the LLM to process.
- **Text-to-Speech (TTS): ElevenLabs**

    - **Justification:** ElevenLabs, specifically the eleven_turbo_v2 model, was chosen for its ability to generate highly realistic and natural-sounding human voices. A custom voice was selected for Dr. Freddy to create a consistent and empathetic persona, enhancing the user's sense of human-like interaction.
- **User Interface (UI): Gradio**
    - **Justification:** Gradio was selected for its simplicity and speed in creating interactive web UIs for machine learning models. It provides ready-made components for audio recording, image uploading, and text/audio display, which significantly accelerated the development of the user-facing application.

Chapter 3: System Design and Methodology

3.1. System Architecture

The architecture of Dr. Freddy is designed as a modular, pipeline-based system where data flows sequentially through different processing stages, from user input to final output. The entire process is orchestrated by a central Python application logic that integrates various external AI services via API calls.

The overall workflow can be summarized in the following steps:

1. **User Input:** The user interacts with the Gradio interface to record their medical query as an audio file and upload a relevant image (e.g., a photo of a skin rash).
2. **Audio Processing:** The recorded audio is sent to the Groq API, which uses the OpenAI Whisper model to transcribe the speech into text.
3. **Image Processing:** The uploaded image is converted into a Base64 encoded string, a format suitable for transmission over an API.
4. **Multimodal Analysis:** The transcribed text and the encoded image are sent to the Meta-Llama model hosted on Groq. A carefully crafted system prompt instructs the model to act as "Dr. Freddy" and analyze the inputs from a medical perspective.
5. **Response Generation:** The LLM processes the combined inputs and generates a textual response in a single, concise paragraph, adhering to the persona defined in the prompt.
6. **Voice Synthesis:** The generated text is sent to the ElevenLabs API, which converts it into a natural-sounding audio file using the pre-selected voice for Dr. Freddy.
7. **Output Display:** The Gradio interface simultaneously displays the textual response and auto-plays the generated audio file, providing a comprehensive, multimodal output to the user.

This architecture is visually represented in the flowchart below.

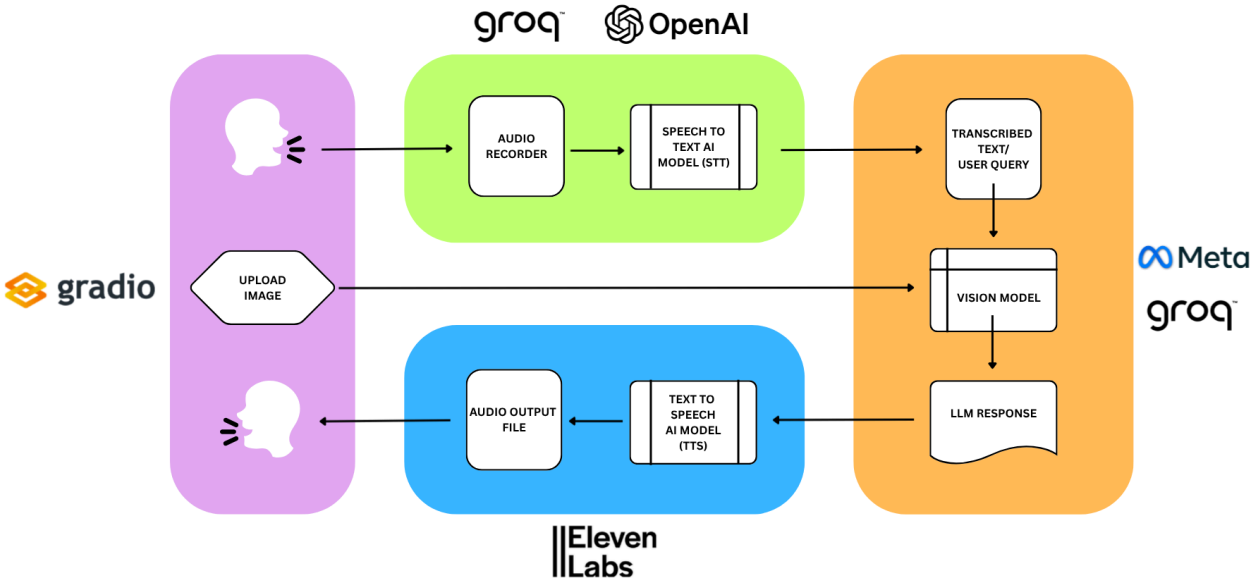

Figure 3.1: System Architecture Flowchart

3.2. Core Logic and Persona Implementation

The "brain" of Dr. Freddy is driven by the Meta-Llama model, but its behavior is strictly guided by a custom system prompt embedded in the gradio_app.py file. This prompt is a critical component of the methodology, as it coerces the general-purpose LLM to adopt a specific, professional persona. The prompt instructs the model to:

- Adopt the name "Dr. Freddy."
- Always respond in a professional, doctor-like manner.
- Analyze the provided image and text from a medical standpoint.
- Provide potential remedies or suggestions.
- Format the response as a single, concise paragraph, avoiding lists or special characters.
- Maintain a natural, conversational tone (e.g., saying "With what I see..." instead of "The image shows...").
- Gracefully handle non-medical questions.

This method of "prompt engineering" is a key part of the implementation, ensuring that the user experience is consistent and aligned with the project's goals without requiring fine-tuning of the model itself.

Chapter 4: Implementation and Results

4.1. Implementation Details

The project was implemented in Python, with the code organized into a modular structure to ensure clarity and maintainability. The core logic is divided into several scripts, each responsible for a specific part of the workflow:

- **gradio_app.py**: This is the main script that orchestrates the entire application. It builds the Gradio user interface, captures user inputs, and calls the other modules in sequence to process the data and generate the final output. It also contains the critical system prompt that defines Dr. Freddy's persona.
- **brain_setup.py**: This module handles the interaction with the core AI model. It contains functions to encode the user's image into Base64 format and to send the multimodal (text + image) payload to the Meta-Llama model via the Groq API.
- **user_voice.py**: This script is responsible for handling the user's voice input. It utilizes the speech_recognition library to capture audio from the microphone and then sends this audio to the Groq API for transcription using the OpenAI Whisper model.
- **doctor_voice.py**: This module completes the interaction loop by converting Dr. Freddy's textual response back into speech. It interfaces with the ElevenLabs API, sending the text and receiving an audio stream which is then played back to the user.
- **Pipfile and Pipfile.lock**: These files manage the project's dependencies, ensuring that the correct versions of all libraries (such as gradio, groq, and elevenlabs) are used, making the project reproducible.

API keys for Groq and ElevenLabs were managed securely as environment variables, which is a best practice for handling sensitive credentials.

4.2. Results & User Interface

The final implementation is a fully functional web application that successfully meets all the project objectives. The user is presented with a clean, intuitive interface designed with Gradio. As shown in the screenshot below, the UI provides clear components for uploading an image, recording an audio query, and submitting the information for analysis.

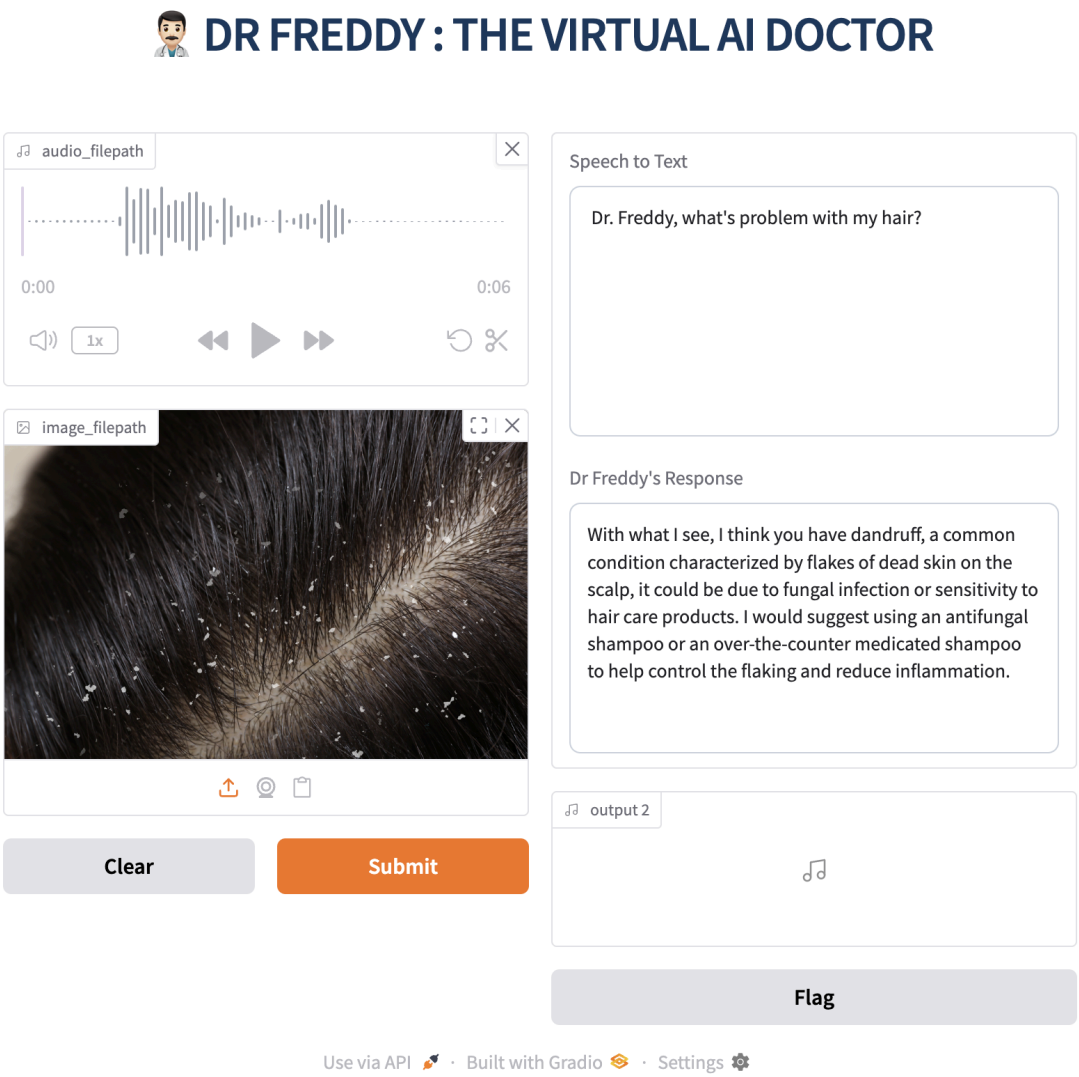

Figure 4.1: Dr. Freddy User Interface

Upon submission, the application processes the inputs and, after a short processing time, displays the text response from Dr. Freddy while simultaneously auto-playing the corresponding audio. This dual-modality output creates a highly engaging and accessible user experience, successfully mimicking a preliminary conversation with a medical professional.

4.3. Testing

Testing was conducted iteratively throughout the development process. This involved:

- **Unit Testing**: Each module was tested independently to ensure its core function worked as expected (e.g., testing the voice transcription in isolation).
- **Integration Testing**: The full pipeline was tested by running the main gradio_app.py script to ensure that data flowed correctly between the different modules and API services.
- **User Acceptance Testing**: The application was tested from a user's perspective to identify any issues with the user interface, clarity of instructions, or the quality of the generated responses. For example, various medical and non-medical queries were posed to Dr. Freddy to test the robustness of its persona and its ability to handle out-of-scope questions gracefully.

Chapter 5: Conclusion and Future Work

5.1. Conclusion

This project successfully demonstrated the development of "Dr. Freddy," a multimodal AI conversational agent for preliminary medical interaction. By integrating state-of-the-art technologies for speech-to-text, multimodal language processing, and text-to-speech, the project achieved its goal of creating an interactive and intuitive tool for users to articulate their health concerns. The application's ability to process both voice and image inputs represents a significant step towards more comprehensive and human-like AI in healthcare. The modular architecture and reliance on robust APIs provide a solid foundation for a scalable and maintainable system. Ultimately, Dr. Freddy serves as a powerful proof-of-concept for the potential of AI to enhance patient engagement and accessibility in the medical field.

5.2. Challenges Faced

During development, a notable challenge was the deprecation of certain AI models available on the Groq platform. This required a careful review of the official Groq documentation to identify the latest, most capable models (such as meta-llama/llama-4-scout-17b-16e-instruct) and update the API calls accordingly. This experience underscored the importance of staying current with the rapidly evolving AI landscape and the critical role that official documentation plays in successful project implementation.

5.3. Future Scope

While Dr. Freddy is a successful prototype, there are numerous avenues for future enhancement:

- **Model Fine-Tuning**: To improve medical accuracy, the underlying LLM could be fine-tuned on a specialized dataset of medical literature and anonymized patient dialogues.
- **Expanded Knowledge Base**: The system could be integrated with verified medical databases (like WebMD or Mayo Clinic) to provide more detailed and referenced information.
- **Session History**: Implementing a feature to save and recall past conversations would allow for more personalized and context-aware follow-up interactions.
- **Deployment and Scalability**: The application could be containerized using Docker and deployed to a cloud platform (like AWS or Google Cloud) to ensure high availability and scalability for a larger user base.
- **Integration with Wearable Devices**: Future versions could potentially connect with health-tracking devices to incorporate real-time biometric data into its analysis.

References

- Malik, P., Pathania, M., & Rathaur, V. K. (2019). Artificial Intelligence in Medicine. *Journal of Family Medicine and Primary Care*, *8*(6), 1937–1941. https://pmc.ncbi.nlm.nih.gov/articles/PMC8285156/
- Official documentation for Gradio, Groq, and ElevenLabs.
- Various educational programming tutorials on YouTube for implementation and debugging assistance.