Project: Building a Voice Assistant Interface for Audio-based LLMs

The emergence of large language models (LLMs), such as OpenAI's GPT-3, has revolutionized natural language processing tasks, enabling various applications in text generation and understanding. One promising application of LLMs is a voice assistant that can interact with users with high-quality synthesized speech.

In this project, we built an end-to-end interface that takes in audio input from a user, transcribes it with the OpenAI Whisper model, feeds the transcribed text into pre-trained large language models (*e.g. a ChatGPT-like model)*, and then renders the response using the Bark model (a generative text-to-speech).

This system was built on top of FastAPI and was deployed to a virtual GPU server for inference acceleration and data collection. To evaluate this system, we recruit participants from the Erdo Institute and/or our friends. They were asked to interact with the voice assistant and then rate the response (like/dislike). We collected the preliminary data to access the usability of this system and potential issues with the current system for future development.