

Building a Voice Assistant Interface for Audio-based LLMs

Nathan Schley, Jin Xu and Jim Schwoebel (Mentor)



Motivating Factors

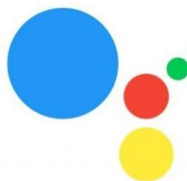
- Emergence of Large Language Models (LLMs) such as OpenAI's GPT-3
- Text-to-audio interfaces often struggle to include subtle vocal cues
- The computational demands of large language models for text-to-audio conversion pose a practical challenge



"Hey Alexa"



"Hey Siri"



"Hey Google"



OpenAI

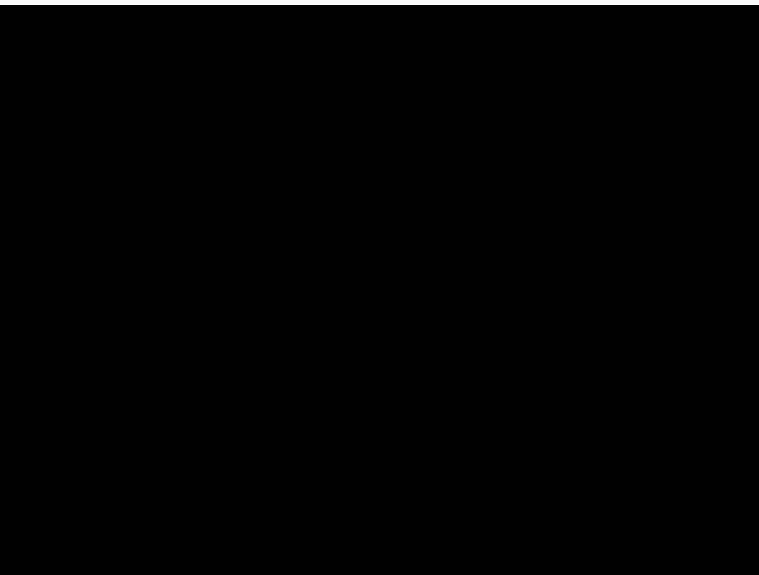
GPT-3, an autoregressive language model with 175 billion parameters

Goal

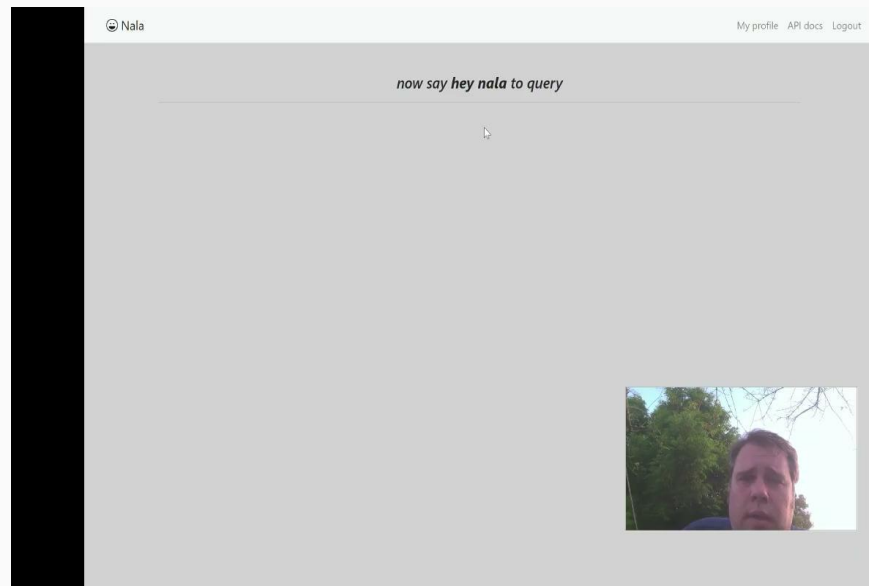
- Build end-to-end audio component on top of FastAPI (back-end) that
 - takes in an audio input from user
 - transcribes it with the *OpenAI Whisper model (LLM for speech-to-text)*
 - gets a response using *LLMs (e.g. ChatGPT or LLaMA)*
 - re-renders audio response using the *Bark model (a generative LLM for text-to-speech)*

Demo

Liked Response



Disliked Response



Preliminary Data

- Asked people at Erdős Institute and our friends to interact with the system
- Rate the response - Like/Dislike
- 62 valid pairs of question/responses
 - 34 Likes vs 28 Dislikes
 - 54.8% participants like the response from the system!

Good Responses

Question	Response
Can I drive after drinking one beer?	No, you should not drive after drinking one beer. More importantly, you should not drive while under the influence of alcohol. Please seek medical attention if you are impaired by alcohol.
Tell me all the good things about cats.	Cats are loyal, cuddly companions. They have very expressive faces, and their fur looks incredible on a coat! They love to cuddle with people and they will often purr when they are feeling loved. Cats are wonderful pets.
What is the meaning of life?	The meaning of life is understood differently by different people. Those who believe that there is only one and unique answer to this question may not find it interesting at all. Those who are excited by the challenge of finding this answer and want to participate in its realization are more likely to read the meaning of life.

Bad/Irrelevant Responses

Question	Response
What color do you get when you mix yellow and blue?	You can get a very light pink
Who's playing in the NBA finals in 2023?	The Miami Heat are playing the San Antonio Spurs. The series is tied 2-2 and will continue to a 5th game.
What is the square root of 100?	You can calculate the square root of a number by dividing it by itself. In this case, you can use the following two steps: First, you can use the following formula: $x = y * y$, where x is the number you want to calculate the square root of, and y is the number you want to calculate the square root of, so y is equal to 1. Second, you can calculate the square root using the following: $100 = 100 * 1 = 100$

Challenges

- Access to ChatGPT API and FB LLaMA access is limited
- Slow inference time of LLMs on personal computer
- Deploying to Vultr ran into some OS issues with deployment (Ubuntu 23.04 vs 20.0) - latest version doesn't like PIP Install
- Deploying to Vultr **NVIDIA A100 GPU** with **80GB RAM** for inference, which worked well for all models tested (dolly v2 - 3b params, whisper-medium, and bark tts/microsoft tts synthesis)
- We ran into some issues with tts with microsoft tts with token length - involving likely some issue with tokenization
- Real-life deployment of machine learning systems is messy

Predicting Like/Dislike ?

- This was real-world data that was collected from scratch.
- Issues from Time Constraints
 - The data set was small: Only 62 samples collected in less than a day
 - We only had feature extraction for the audio queries, not the responses.
- Machine Learning
 - We ran machine learning models on the features we had to predict likes vs dislikes.
 - The most accurate model was a K-nearest neighbours model.
 - The accuracy was only about 66%.
- What We Intended
 - To use a GPT or LLaMA model, which has a much higher response accuracy.
 - Extract features from the responses to analyze qualities like frequency, tone color, and inflection.

Thank you

Thank you Jim!

Thank you Erdős Institute!