Spring Semester 2024

LLMs for human-robot interaction autonomous conversations

Mortadha Abderrahim
Supervisor: Daniel Carnieto Tozadore

MOTIVATION

The rapid advancement of large language models (LLMs) has opened up new possibilities in various applications, including intelligent tutoring systems in education. Social robots, equipped with speech recognition and text-to-speech systems, can potentially serve as effective educational tools. This project explores the integration of LLMs into social robots to enhance educational interactions, aiming to understand if fine-tuning LLMs on educational data improves their performance and how these models affect human speech behaviors during interactions

METHODS

The system design involved integrating an AVSR-LLM pipeline, which includes audio and video recording, transcription, LLM response generation, and robot text-to-speech. To enhance the interaction flow, a Voice Activity Detection (VAD) mechanism was implemented to stop recording when users stop speaking. Three large language models were evaluated: StableLM Zephyr 3B, Phi-3-Mini-128K-Instruct, and Phi-3-Mini-128K-Instruct fine-tuned on educational data. These models were benchmarked using datasets such as MMLU, AI2 Reasoning Challenge, and Hellaswag to assess their performance.

Additionally, a speech feature classifier was developed to distinguish between human-human and human-robot interactions based on speech features.

A pilot study was conducted where participants interacted with the three LLMs on various topics. Data on user preferences and speech behaviors were collected and analyzed to determine the effectiveness of the models.

RESULTS

Benchmark performance indicated that the fine-tuned Phi model outperformed the others on educational benchmarks, demonstrating its effectiveness in educational settings. The pilot study findings revealed that participants preferred the fine-tuned Phi model, which also influenced their speech behaviors to be more human-like. However, technical limitations were identified, such as processing delays due to hardware constraints and the need for more robust transcription technologies to improve interaction quality.

Link to the code repo and report: https://github.com/Mortadha-abderrhim/AVSR_VAD_LLM