MIDA

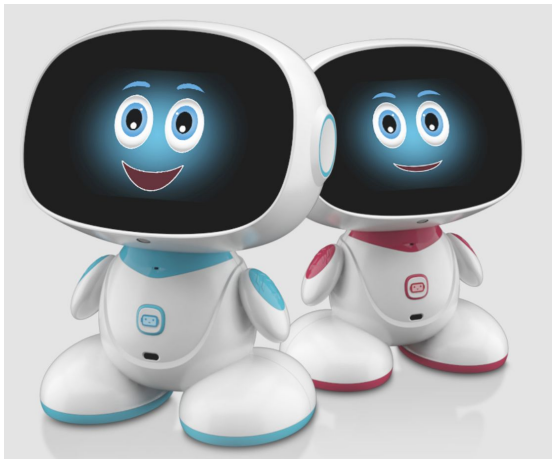
GPT-powered Furhat Robot as a Multilingual Information Desk Assistant

Bogdan Laszlo & Aruna Elentari

January 11, 2024

Why: Great Value in LLM Powered Social Robots

- LLMs such as chatGPT and LLaMA-2 are powerful tools
- Social robots are being used in hotels (Connie at Hilton), hospitals, homes (MISA, aibo), etc
- Powering robots with LLMs is a logical next step



What: Project Goals

Learning:

- Learn how to integrate the Furhat robot with GPT-4

Concrete:

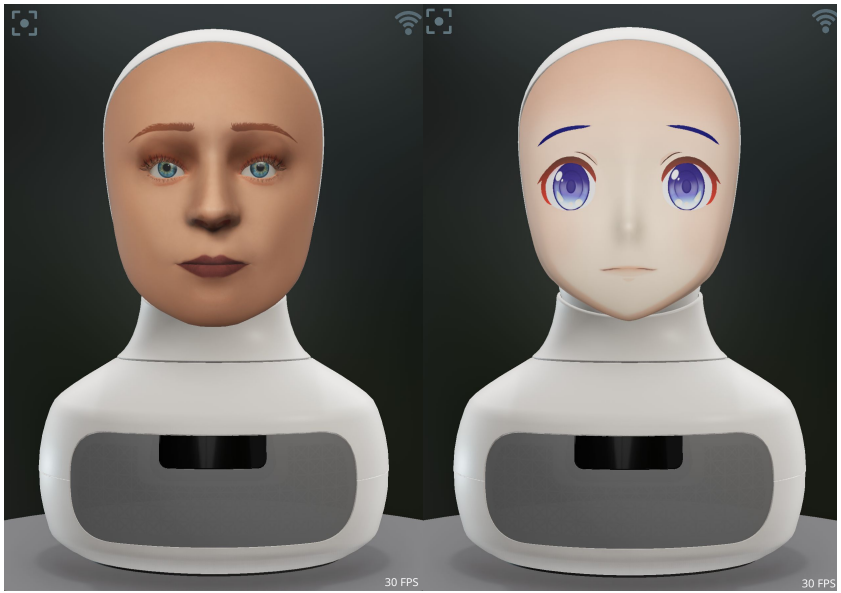
- Demonstrate sample interactions between a tourist and MIDA in multiple languages
- Have MIDA recognize flags of various countries and start a conversation in the language of that country

Why Furhat: expressive face, onboard audio/video sensors, 40+ languages, virtual, smooth head movements

Why GPT: easy to use, Furhat Robotics had available code for GPT integration

How: Furhat Dialogues

- Languages: English, Russian, Swedish, Spanish
- Setting: airports in Barcelona and Gothenburg, a visitor center in Gothenburg, Red Square in Moscow
- Using Blockly for crafting dialogues
- Insights
  - Use simple target words for questions
  - Break the dialogue into separate states
  - Take into account silence, users leaving
  - Challenges: Russian, mishearing users



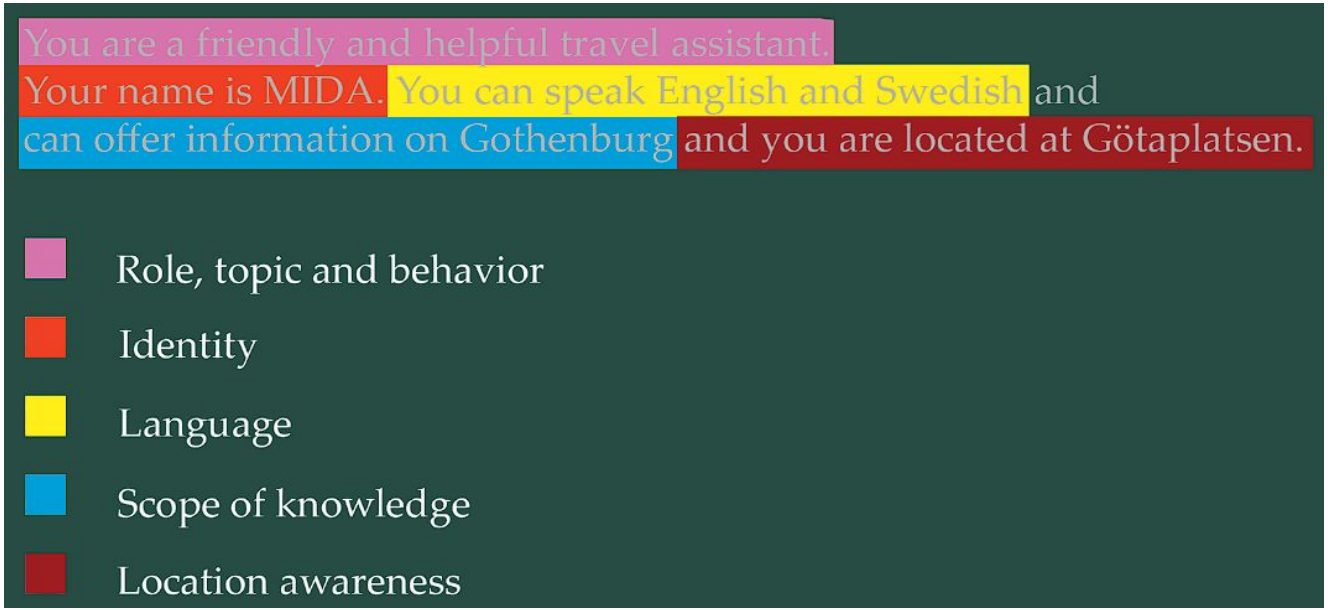
MIDA: Furhat + GPT

- Implemented a curl request function in Kotlin in order to send prompts and receive responses with our temperature setting
- Used prompt engineering in creating both MIDA's custom initialization instructions and in wrapping the user queries
- Implemented targeted conversation context for GPT by injecting into the dynamic prompts information from the past queries within the conversation alongside their purpose

Custom Instructions and Prompt Engineering

- Each MIDA GPT instance is initialised with a custom instruction

The instructions can be split into 5 parts based on their purpose



Mixed Language Prompt Engineering

In order to achieve the best results, even for Non-English prompts, English language prompt engineering was used

Language	Mixed	Non-English	Change%
Russian	152	190	-20
Swedish	71	114	-37
Spanish	62	83	-25

[custom instruction][user prompt]
Please don't use more than 50 tokens.
USE 50 TOKENS OR LESS PLEASE.

In-prompt output condition specification

In-prompt condition assertion

Example of mixed user prompt composition

Challenges

- API unpredictability, as it could send timeout errors with no prior warning or no apparent reason for doing so
- Long latency times for API calls, where depending on the complexity, length and language of the prompt, the API would take visibly long to produce a response

MIDA and Flags

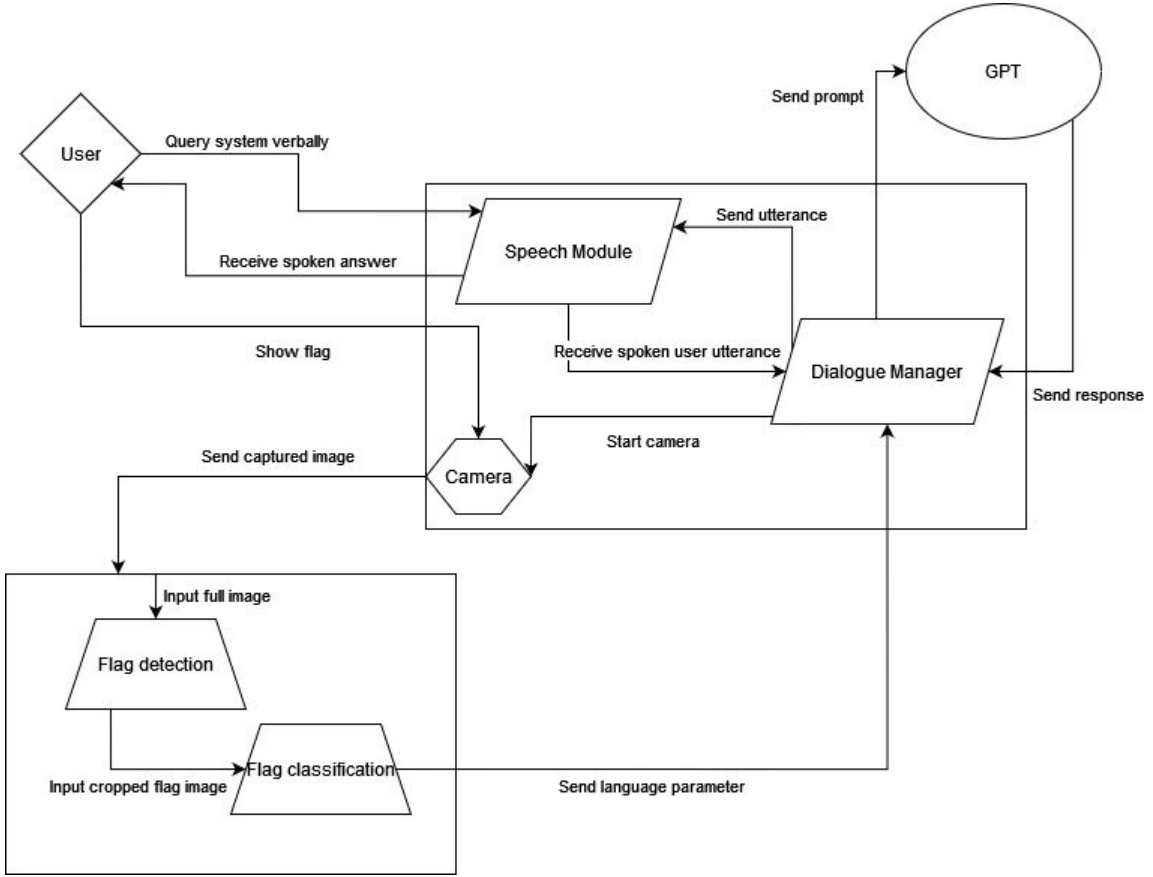
- In order to showcase Furhat’s reactivity to the world around it we have implemented a language and scenario change conditioned by flag detection
- When asking Furhat to change its language, Furhat will activate its in-build camera module, capture a snapshot of its surroundings, detect and classify the flag in order to extract the desired conversation language

Dataset

- The dataset was manually collected by selecting freely available images of flags in various environments from the internet, but also by taking photos of flags in environments similar to those MIDA would operate in (i.e. low light indoor environments)
- The dataset was manually annotated, thus providing bounding boxes, object types and assigning languages

Model Training

- MIDA used 2 models chained together, both based on the ResNet50 architecture.
- The first model detects the objects of interest from the snapshot and passes their cropped image to the classification model.
- The classification model then based on the flag's colours and shape classifies them into their respective languages accordingly



Challenges

- Assembly of such a dataset takes time, which scales with the flag diversity, which is why only 3 Non-English languages are implemented
- While the object detection model behaves adequately, the flag classification one suffers from the low amount of data in the dataset (below 1000 flags), not allowing it to adapt to camera resolutions and lighting conditions

Language	Precision	Recall	F1-Score	Support
Unknown	0.38	0.16	0.22	76
Spanish	0.13	0.15	0.14	34
Russian	0.14	0.26	0.19	38
Swedish	0.21	0.25	0.23	40

Demo

- Furhat dialogue in Spanish at the Barcelona airport (recording)
- MIDA at the visitor center in Gothenburg
- MIDA recognizing a flag and switching to the language of the flag

Conclusion

- Embodied agents powered with LLMs can have great value as assistants in various settings
- Open-source LLMs can be used, either by connecting to internet servers or perhaps even locally running models with updateable knowledge base
- One can leverage GPT's multimodal capabilities in helping users, particularly in areas of high cultural diversity