MIDA: GPT-powered Furhat Robot as a Multilingual Information Desk Assistant

Aruna Elentari **Bogdan Laszlo**
LT2318 Artificial Intelligence: Cognitive Systems
Gothenburg University

Abstract

Our project involved integrating the Furhat robot with GPT-4 to develop a Multilingual Information Desk Assistant (MIDA) that could answer tourists' questions in a friendly and succinct manner. This system should be able to handle complete task-oriented dialogues in English, Spanish, Swedish and Russian, as well as being able to detect flags and recognize their associated language. Keywords: Furhat, GPT, Multilingual Information Desk Assistant (MIDA)

1 Introduction (AE)

The field of artificial intelligence (AI) is currently experiencing a lot of breakthroughs, primarily driven by large language models (LLMs), such as chatGPT and LLaMA-2, both of which are transformers (Naveed et al., 2023). These models receive large amounts of varied textual data as part of their training, and are able to generate compelling text based on input prompts. One of the latest LLMs, GPT-4, is a multimodal model which receives image and text inputs and produces text outputs (et al., 2023). In the last few years, we have also seen an increased interest in the use of social robots in various settings, such as hospitals and hotels (Vishwakarma et al., 2024). All these developments inspired us to explore powering social robots with LLMs to develop a friendly assistant that can be used in tourism.

Our project involves integrating the Furhat robot with GPT-4 to develop a Multilingual Information Desk Assistant (MIDA) for use at a tourist information desk. We chose to work with the Furhat robot since it was included in the course curriculum and has several advantages. There is no need to interact with the physical robot, although one is available at the University of Gothenburg. Instead, one can interact directly with the virtual Furhat by downloading the software development kit (SDK) on your computer. The dialogues with Furhat can be written in Blockly, a visual programming editor that is accessible to people with little or no programming background. Apart from being easy to use, Furhat has a highly expressive face, human-like head movements, onboard sensors for audio and visual perception, and over 40 supported spoken languages.

A somewhat similar experiment was done by (Abbo and Belpaeme, 2023), where they integrated GPT-4 with Furhat to create six dialogues that took place in five different environments: a lab, a kitchen, the home entrance, a bathroom, and a bedroom. Their system consisted of four components: the dialogue processing, which received auditory input from the user, the frame processing, which received video input from the user, the dialogue manager, that received inputs from the first two components, and relayed that information to the GPT-4 component, which in turn sent outputs to the dialogue manager. The authors concluded that integrating the visual information from the video feed with the audio information allowed for a richer, contextually aware dialogue between the human user and the robot, and that this is a promising area to explore in human-robot interaction.

When using chatGPT as a multilingual tourist assistant, one needs to take into account the fact that the bulk of information that was fed into LLMs during training was in English and a few other dominant languages on the internet. When (Kolar and Kumar, 2023) used chatGPT as a tourist assistant to translate travel related text from English into Hindi, Telugu and Kannada languages, they discovered that Hindi translations were most accurate and fluent, while the other two lagged behind. This is not surprising given that Hindi is highly prevalent in India, and it is one of

the top ten most spoken languages in the world. There also needs to be a robust dialogue evaluation methodology.

As part of their experiment involving the GPT-integrated Furhat robot, (Abbo and Belpaeme, 2023) highlighted the main criteria for evaluating interactions between humans and robot assistants, such as mutual understanding and common ground, smoothness of interaction, active listening, turn-skipping and trustworthiness. The assistant needs to remember certain aspects about the user that are relevant to the conversation, be able to stick to the subject, and handle missing, inconsistent or irrelevant information in a graceful manner.

We took many of these measures into account when designing our system. Our project encompasses not only GPT integration with Furhat, but also using Furhat's and GPT's multilingual capabilities to have dialogues in multiple languages. In section 2, we present the concrete goals of this project. In section 3, we describe our methodology in detail. In section 4, we present our results, and in sections 5 and 6, we discuss our results and share our conclusions.

2 Concrete Goals (AE)

Apart from gaining knowledge and skills, we wanted to achieve demonstrable goals. The main concrete goal of the project was to demonstrate sample interactions between a tourist and MIDA in multiple languages. We wanted to reproduce a typical interaction between a tourist and a person working at an information desk at an airport or a visitor center. The second goal of the project is to have MIDA recognize flags of various countries and start a conversation in the language of that country.

First, we wanted to develop a minimum viable product in the form of dialogues with Furhat alone. We designed four dialogues in four languages: English, Swedish, Russian and Spanish. This allowed us to test Furhat's dialogue capabilities more fully and cover three of the most spoken languages in the world. Next, we wanted to integrate Furhat with GPT-4 to create MIDA, which would be able to answer users' travel related questions in any of the four aforementioned languages. Last, we combined MIDA with an object detection model that was trained specifically on flags, so that it could recognize the flag that is being shown and converse in the official language of the flag's country, which in our case was one of the four chosen languages.

3 Materials and Methods (BL)

3.1 Furhat Dialogues

All dialogues start in English, and then three of them switch to the appropriate language upon user request. To stay within the scope of the project, all interactions are between a single user and Furhat. Apart from helping us learn about Furhat's capabilities, these dialogues act as a reference for an ideal interaction, which can be used to test the quality of dialogues with GPT-integrated MIDA. All dialogues were first prototyped in Blockly, and later on reimplemented in Kotlin, with questions including target words to make it easier for Furhat to respond to them. All the voices are synthesized by Amazon Polly.

The first scripted dialogue is in Spanish and takes place at the Barcelona airport. The Furhat character is Isabel (adult), the Spanish voice is by Lucia (es-ES). It was necessary to switch to Lucia's voice to have a correct Spanish pronunciation. The second dialogue is in Swedish and occurs at the Gothenburg airport. The character is Jane (adult), the Swedish voice is by Elin (Neural, sv-SE). The third dialogue is in English and takes place in Gothenburg's visitor center. The character is anime (legacy), the English voice is by Aria (Neural, en-NZ). While in the previous two dialogues Furhat has a human face, in this and the following dialogue, we used an anime face. While we have not explored the importance of human faces in robots in this project, we were interested in seeing the effect of an anime face during the presentation. The fourth dialogue is in Russian and takes place in the Red Square in Moscow where the tourist asks for sightseeing and restaurant recommendations from Furhat. The characters are anime (legacy), the English accent is by Aria, and the Russian is by Tatyana (ru-RU).

3.2 MIDA: GPT-powered Furhat

In order to enforce MIDA's role within conversation with user agents, a custom instruction was passed through the GPT API when being initialized. This instruction was divided into 5 parts in order to properly configure the GPT instance for the role it's going to fill. The instruction is written in the main language that GPT is going to interact

with the user in, in order to make sure that it will start in that language from the onset of the conversation. The custom instruction's parts are as follows:

- Role, topic, and behavior defining "You are a friendly and helpful travel assistant," which sets the tone, choice of words, and topic for any prompts MIDA will receive.
- Identity "Your name is MIDA," which makes sure that GPT knows what name to use, recognize it's being addressed, and so on.
- Language "You can speak English and Swedish," this way GPT will be preemptively conditioned to accept input in those languages, but also will know how to handle words, locations, events, and so on in those specific languages.
- Scope of knowledge "offer information on Gothenburg," which limits the boundaries of the information MIDA can provide, which leads to an increase in precision and relevance in its answers.
- Location awareness "you are located at Götaplatsen," this allows GPT to offer relevant answers when directing the users towards various points of interest.

Not only that but the GPT implementation is also able to leverage previous dialogue context within the same conversation. In order to minimize any potential misunderstandings and erroneous decisions on its part, the accessible context is curated, only certain elements with relevant information is included in it, such as suggested points of interest for dynamically giving directions towards.

3.3 Flag Detection and Classification

For flag detection and classification, two ResNet50 models are chained together. This system consists of a pre-trained Faster R-CNN model with a ResNet-50-FPN backbone (Ren et al., 2016) which is fine-tuned to properly detect flag objects. Once those objects are detected and their bounding boxes are generated based on a single snapshot, the input image is then cropped, resized and input into a pre-trained ResNet50 model, which is fine-tuned for classifying flags based on the language associated with them. The dataset was assembled specifically for this research using both freely available images from the internet, but also taking pictures of flags displayed on phones mimicking the various lighting conditions that MIDA may encounter. The annotations were done manually, tracing the bounding boxes, labeling the objects based on object type (flag or not) and language, by the end the dataset amounted to a total of 1038 objects. The end result of this system being the information regarding what language MIDA should use in its voice input and output capabilities.

These systems are then integrated into the Furhat platform, the flag detection and classification system acting as an external processing server, with its own logic separate from Furhat's code. It is activated when prompted by the user within Furhat's dialogue states, which triggers its connection to Furhat's in-built camera, allowing it to receive visual information. GPT is used within Furhat's dialogue states for giving dynamic and complex information to user queries, while simple pre-written utterances are used for addressing follow-up questions to the user, acknowledging the user's queries and greetings. This hybrid system ensures that this Furhat system, while complex, manages to be responsive in its actions, an important feature for any human-facing system.

MIDA's general dialogue flow is as such:

- MIDA is manually activated by the user.
- MIDA prompts the user to state their query in English.
- The user can either state the query directly, the conversation remaining in English, or can ask MIDA to change the language (for the sake of the experiment, this leads to simulating different scenarios).
- The user and MIDA have a dialogue with the purpose of providing the user information on the chosen city's attractions and means of transport.

4 Results (BL)

MIDA manages to achieve the desired goals as stated in the introduction, specifically implementing a dynamic GPT-powered chatbot with multilingual and multimodal capabilities. The system was tested in various lighting conditions and multiple languages, exhibiting varying performance.

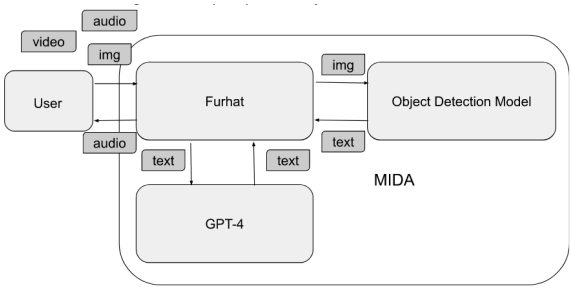

Figure 1: A high level diagram of MIDA. MIDA consists of Furhat, GPT-4, and an object detection model. The user interacts with Furhat by speaking to it and showing it images of flags. It converts the audio signal (user questions) to text and sends it to GPT-4, which sends back the text output (answers). Furhat relays the image of a flag, such as a flag of Sweden, to the object detection model, which in turn outputs the text "Swedish" back to Furhat.

The flag detection and classification system, while not perfect, manages to provide a great performance when detecting flags, as seen during the live Furhat tests. Regarding flag classification, the performance can be considered lackluster, as seen in Table 1, performing better on well lit controlled environment high resolution images and performing worse on images replicating scenarios that MIDA could encounter.

Table 1: Flag Classification Results

Class	Precision	Recall	F1-Score	Support
0	0.38	0.16	0.22	76
1	0.13	0.15	0.14	34
2	0.14	0.26	0.19	38
3	0.21	0.25	0.23	40

Metric	Score
Accuracy	0.1968
Macro Avg	0.21
Weighted Avg	0.25

MIDA also manages to achieve satisfactory performance in all target languages, even when presented with a non-proficient speaker of such languages. Furhat's auditory recognition capabilities offer good results in recognising spoken words, while GPT can understand and produce text in those languages of great quality. One caveat of using multiple languages that can be noticed is that prompt engineering has diminished returns when being done in languages other than English. This problem was mitigated by offering mixed language prompts, where information such as desired token amount and information focusing prompts were specified in English, as it is visible in Table 2.

Table 2: Showcasing the usage of the prompt engineering feature for limiting the tokens used to below 50. The full English control prompt used only 44 tokens

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

5 Discussion (BL)

Our project manages to achieve a working information desk focused automated agent. As per our findings, it manages to perform its tasks using multiple languages and answering questions without being sidetracked. While building on findings of papers such as (Abbo and Belpaeme, 2023) and (Kolar and Kumar, 2023), MIDA manages to address some of their concerns while encountering some of the same challenges as them.

(Abbo and Belpaeme, 2023) implemented a similar Furhat system with more complex capabilities, which suffered from a predisposition to being led off topic by the multitude of information sources it could use. MIDA due to its focused approach does not encounter such a problem due to a mix of using visual information only in a strictly on-demand manner, using dialogue states in order to guide the flow of the conversation rather than allowing a complete free flowing conversation and the use of prompt engineering in order to enforce the length and topic of answers.

One problem that we also encountered was the limitation brought about by the object detection server, which due to our hardware could not be used in real-time, a problem that was partially mitigated by having it capture a single frame on demand, which while not ideal, offered a suitable workaround for the problem. Another similar issue which we have encountered is the GPT API's delay in generating answers. This problem was partially addressed by keeping the amount of tokens low, but did not offer a consistent solution, non-English prompts taking a longer time to generate compared to English ones.

(Kolar and Kumar, 2023) research shows that

GPT does not fare well with languages significantly different than English and with a lower presence in its training corpus, which is the reason why we chose well documented languages which are spoken internationally by a large amount of people (Spanish, Russian) or that are sufficiently similar to English so that it won't pose a problem for it (Swedish). GPT's performance in generating answers was exceptional, providing useful and grammatically correct answers, as judged by native speakers of those languages.

6 Conclusions and Further Work (AE)

We successfully integrated the Furhat robot with GPT-4 to develop MIDA specifically for use at a tourist information desk. MIDA was able to answer travel related questions and provide requested information in the languages of our choice. The next step would be to use open-source LLMs such as LLaMa 2, in combination with Furhat and other social robots. This would allow for more exploration in designing the best combination of robots and LLMs for specific purposes, which extend beyond tourism.

One needs to take into account various ethical considerations when designing an actual product to be used by many people. These include potential discrimination based on language or ethnicity, being culturally or politically unaware, sending tourists only to locations which paid money to be featured, or profiling people and reporting the data for other uses (government, police or just selling on the data market). One also needs to be particularly attuned to potential time delays in processing data.

Possible future developments of this Furhat-based system could consist of increased multimodal capabilities, leveraging LLMs such as Gemini or GPT. One such implementation could be Furhat's ability to receive visual input in the form of images of maps or simply images of the point of interest and request information on them such as planning a route using a particular means of transport or simply obtaining opening time information.

While working on this project, we were reminded of the amazing abilities we humans have in interacting with one another and using many nonverbal cues in communication. In a real-life tourist situation, there probably needs to be a human working next to a robot assistant powered by LLMs, at least initially. There are a lot of opportunities in exploring this particular area of robotics and AI, and we hope to see more multi-modal LLM powered robots of various shapes that can be great assistants for humans in the near future.

References

Giulio Antonio Abbo and Tony Belpaeme. 2023. I was blind but now i see: Implementing vision-enabled dialogue in social robots.

OpenAI et al. 2023. Gpt-4 technical report.

Sanjana Kolar and Rohit Kumar. 2023. Multilingual tourist assistance using chatgpt: Comparing capabilities in hindi, telugu, and kannada.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Laxmi Pandit Vishwakarma, Rajesh Kr Singh, Ruchi Mishra, Denizhan Demirkol, and Tugrul Daim. 2024. The adoption of social robots in service operations: A comprehensive review. *Technology in Society*, 76:102441.