ABSTRACT

Candidate: *Marco Pardini*

Title: *Development of a Service-oriented Cognitive System for Social Robotics orchestrated by a Large Language Model*

Master's degree in Artificial Intelligence and Data Engineering

Supervisors:

1) Prof. Mario Giovanni Cosimo Antonio Cimino

2) Prof. Federico Andrea Galatolo

3) Prof. Lorenzo Cominelli

With the growing demand for increasingly competent robotic systems for social and emotional human-robot interactions, there arises a necessity for more sophisticated means of control.

The need for advanced cognitive architectures in humanoid robots arises from their role in interacting with humans to solve various tasks, such as assisting with daily activities, providing medical support, and supporting educational endeavours. These interactions require robots to understand and align with human behaviours, emotions, and intentions in a seamless and socially appropriate manner.

In the literature, cognitive architectures have historically relied on assumptions about the human mental model, often employing rule-based systems or Markov Decision Processes (MDPs). Alternatively, some learned human mental models through interaction using reinforcement learning approaches or a combination of both. These methods have inherent limitations, as they struggle to generalize across diverse scenarios. [5]

The advent of LLMs has introduced a new avenue for approximating human mental models. This development has sparked an expanding field of research focused on integrating LLMs into cognitive architectures in various ways [2]. Many LLMs have been released recently, with numerous companies striving to develop their own unique versions. Among the most renowned are GPT-4 from OpenAI, Mixtral8x7b from MistralAI, Llama3 70b from Meta, Gemma-2-9b from Google, Phi-3-Mini 3.8b from Microsoft, and many others.

The paper "ChatGPT for Robotics: Design Principles and Model Abilities" [4] published on April 15, 2024, by IEEE, addresses these limitations by investigating how OpenAI's LLM ChatGPT, known for its interactive dialogue capabilities, can be generalized to the domain of robotics. Instead of merely passing a task description to ChatGPT, the paper proposes providing a prompt along with a set of explained APIs and the objective. ChatGPT then calls a subset of these APIs to interact with the given scenario. The authors qualitatively evaluate the proposed architecture across various robotic tasks, including aerial robotics, industrial inspection, curriculum learning, and most importantly, embodied agents. For instance, they demonstrate making a robot navigate to an area of interest by receiving state observations as dialogue text.

However, the limitation with the embodied agent presented in the paper is its inability to express emotions. It is not designed as a social and emotional robot but rather as a task-oriented robot, focusing on functions such as exploring unknown environments or holding dialogues.

Social and emotional robots must adhere to the principles of eBICA (Emotional Biologically Inspired Cognitive Architecture) [3]. At the core of eBICA is the standard cognitive cycle: perceive, understand, generate ideas of possible actions, select an intention consistent with the working scenario, commit the intended action, check the outcome against expectation, and resolve surprises if any.

To assess the capabilities of Emotional Cognitive Architectures, a few metrics have been adopted in the literature to evaluate cognitive architectures of emotional and social robots, such as experiments with large numbers of participants equipped with heart rate and sweat monitors, and by addressing questionnaires.

The solution proposed in my thesis adapts and expands the framework proposed by OpenAI with ChatGPT, addressing its limitations in social and emotional robotics, and surpassing the state of the art by design. It proposes a novel use case for this kind of architecture.

The LLM agent is equipped with a variety of Functions tailored for human-robot interaction: Reasoning, Talk, Emotion, Hear, Look, Memorize, Recall, Recognize, and Stop. The architecture is built as a service-oriented model, where each Function call is managed by a container running on different machines. These services are decoupled and fault tolerant. Communication between the client (Abel) and the services occurs via MQTT.

The division into function calls is also particularly pertinent to the emotional cognitive architecture, as studies demonstrate the functional specificity of the brain not only for basic sensory and motor functions but also for higher-level cognitive functions.

To adhere to the eBICA principle of contextual perception, continuous environmental feedback is automatically injected into the prompt through autonomous Function calls, triggered by contextual changes. This mechanism leverages two processes running parallel with the main architecture: a context process and a hear process.

The context process ensures the embodied agent is fully aware of its surroundings. If the context changes, the LLM is notified by injecting the new context into the prompt, independently of any calls from the architecture. The hear process ensures that the embodied agent is always listening, even when other function calls are active. Everything said is saved into a queue, allowing the agent to address them later.

Regarding the evaluation, due to resource constraints and the unavailability of the robot for experiments, a qualitative analysis to compare the behaviours of different LLMs has been undertaken, employing a novel approach that leverages process mining transition maps.

As proposed by the guidelines outlined in the paper: "A Primer for Conducting Experiments in Human-Robot Interaction" [1], the procedure involves a series of experiments with 20 participants, each testing four different LLMs to compare their performances, conducted in a rigorous manner. The selected LLMs for the comparison are Mixtral8x7b, Llama3 70b, Gemma-2-9b, and Phi-3-Mini 3.8b, chosen for their differences in both parameter counts and underlying architectures.

A predefined, ambiguous scenario is used, requiring each participant to rephrase it in their own words. The objective is to assess which function calls each LLM performs in ambiguous

situations and to understand if different LLMs exhibit varying behaviours, potentially corresponding to different parts of the human brain.

The results shown by the transition maps are quite similar across models, with cycles between reasoning, talking, and hearing present in all. However, a notable difference in function calls is observed: the number of inbound arcs (69) in the'Look' Function of the Llama3 70b model is near double the number of inbound arcs of 'Look' calls of other models (45, 46, 38), making it particularly aware of its surroundings. This situational awareness contributed to its high appreciation among participants, with 18 out of 20 preferring Llama70b, 2 out of 20 preferring Mixtral8x7b, and none preferring Gemma2 or Phi3.

Regarding future developments, this architecture is particularly suited for experimentation, as each module uses a model that can be arbitrarily changed to test new ones as they are released. Additionally, a faster communication protocol could be used, and fine-tuning the parameters could further enhance conversation quality. Incorporating advanced features such as gesture recognition and facial expression production could enrich interactions. Performing psychological tests on different LLMs to determine if their function call choices indicate distinct psychological behaviors would also be valuable. Rigorous evaluations involving heart rate and sweat rate monitors, along with questionnaires, should be conducted to thoroughly assess the architecture's effectiveness.

Bibliography

[1]

Hoffman, Guy, and Xuan Zhao. "A primer for conducting experiments in human–robot interaction." ACM Transactions on Human-Robot Interaction (THRI) 10.1 (2020): 1-31.

[2]

Kim, Callie Y., Christine P. Lee, and Bilge Mutlu. "Understanding large-language model (llm)-powered human-robot interaction." Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024.

[3]

Samsonovich, Alexei V. "Emotional biologically inspired cognitive architecture." Biologically Inspired Cognitive Architectures 6 (2013): 109-125.

[4]

Vemprala, Sai H., et al. "Chatgpt for robotics: Design principles and model abilities." IEEE Access (2024).

[5]

Verma, Mudit, Siddhant Bhambri, and Subbarao Kambhampati. "Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion?." Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024.