Exploring LLM in Intention Modeling for Human-Robot Collaboration

Sikai Li[1], Run Peng[1], Yinpei Dai[1], Jenny Lee[1], Joyce Chai[1]

I. Introduction

Humans develop Theory of Mind (ToM) at a young age - the ability to understand that others have intents, beliefs, knowledge, skills, etc. that may differ from our own [2]. Modeling others' mental states plays an important role in human-human communication and collaborative tasks. As a new generation of cognitive robots start to enter our lives, it's important for these robots to have similar ToM abilities in order to effectively collaborate with humans. While there is an increasing amount of work in ToM modeling for collaborative tasks in human-agent collaboration, most of the works were situated in a simulated environment [1], [6]. In this work, we take an initial step towards ToM modeling powered by large language models GPT-4 [5] in human-robot communication and collaboration. In particular, we applied prompt engineering in a one-shot setting to empower the robot the ability to infer human's intention and generate corresponding responses.

II. System Overview

As shown in Figure 1, our TIAGo robot and its human partner collaboratively prepare two dishes with specific ingredients determined by human instructions in a real world scenario. There is a table positioned centrally between the robot and the human which has a variety of items and a plate on each side. The human and the robot each can reach for objects close to their own side, but cannot reach objects on the other side.

Our HRC system applies Grounded-SAM model [3], [4] with RGB-D images from TIAGo to localize all the objects on the table. Then the ToM model empowered by GPT-4 processes human utterances (red chatbox in Fig. 1) and the current state (e.g., objects' positions; marked as blue) of the environment. Leveraging GPT-4's robust reasoning capabilities, the model can spontaneously reason about the differences between the goal and the current state, infer what the human is going to do next and what the human may need (dotted bubble in Fig. 1), and then release actions (highlighted in green) to the robot. In addition, GPT-4 makes open language communication possible, where human speech is not limited to specific commands like "give me the apple", but more natural daily communication. For the example shown in Fig. 1, after the human tells the robot to prepare the first dish, the reasoning result shows that the human needs yellow mustard but it is on the robot's side. The

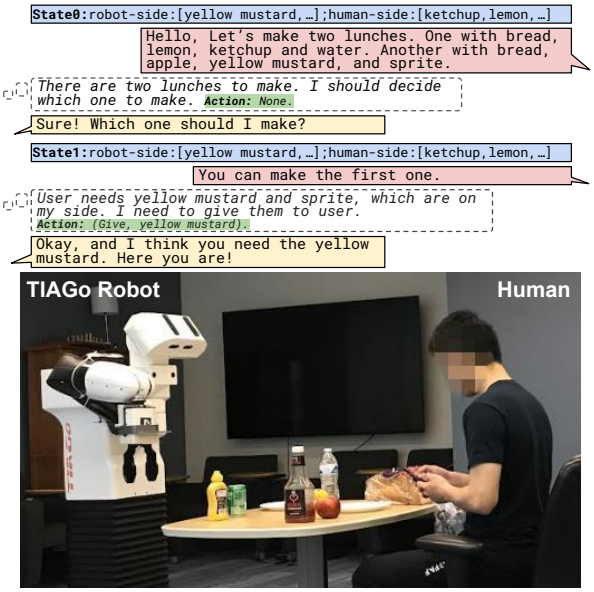

Fig. 1. GPT-4 empowered ToM model infers human intentions, releases actions, and generates corresponding responses to robot spontaneously.

robot predicts that the human's intention is to get the yellow mustard and therefore executes actions to pick up the yellow mustard and hand it to the human.

We conducted preliminary experiments to study the efficacy of our TIAGo robot powered by GPT-4 in a simplified meal preparation task. While the method is promising, our robot still faces many challenges, particularly in sensing and manipulation. Our future work will conduct more comprehensive studies and examine how LLMs may help mitigate these difficulties with the help of human communication.

[1]Authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, MI 48105, USA. skevinci@umich.edu

References

[1] Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[2] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind, 2017.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[5] OpenAI. Gpt-4 technical report, 2023.

[6] Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, 2023.