

Comparing Apples to Oranges: LLM-powered Multimodal Intention Prediction in an Object Categorization Task

Hassan Ali, Philipp Allgeuer and Stefan Wermter¹

Abstract—Intention-based Human-Robot Interaction (HRI) systems allow robots to perceive and interpret user actions to proactively interact with humans and adapt to their behavior. Therefore, intention prediction is pivotal in creating a natural interactive collaboration between humans and robots. In this paper, we examine the use of Large Language Models (LLMs) for inferring human intention during a collaborative object categorization task with a physical robot. We introduce a hierarchical approach for interpreting user non-verbal cues, like hand gestures, body poses, and facial expressions and combining them with environment states and user verbal cues captured using an existing Automatic Speech Recognition (ASR) system. Our evaluation demonstrates the potential of LLMs to interpret non-verbal cues and to combine them with their context-understanding capabilities and real-world knowledge to support intention prediction during human-robot interaction.

I. INTRODUCTION

With the rapid advancements in AI, humanoid robots are set to play increasingly prominent roles in human societies, spanning diverse sectors like manufacturing, healthcare, and domestic environments. This technological shift is expected to foster a natural and intuitive interaction between humans and their robotic collaborators, hence resembling human-human interaction rather than just relying on explicit commands [1]. Intention prediction is fundamental in HRI and involves perceiving subtle cues from human actions, gestures, and speech to infer underlying goals [2]; therefore, intention prediction is key for enabling effective and seamless communication. Nevertheless, interpreting and predicting human actions is challenging due to the multimodal nature of human behavior, being complex and driven by situational context. Also, human intentions are often conveyed with nuanced non-verbal cues which are challenging to interpret.

Typically, intention-based systems require collecting a vast amount of data, which is both impractical and restrictive, limiting the systems' applicability to specific use cases or a particular set of modalities and/or setups [3]. Recent research with LLMs in robotics have showcased their abstract reasoning capabilities going beyond mere language applications. Besides their proficiency in text comprehension and generation, LLMs have exhibited signs of notable inferential skills, extensive real-world knowledge and robust reasoning capabilities [4]. Additionally, they are performant at capturing context, making them a candidate for advancing intention

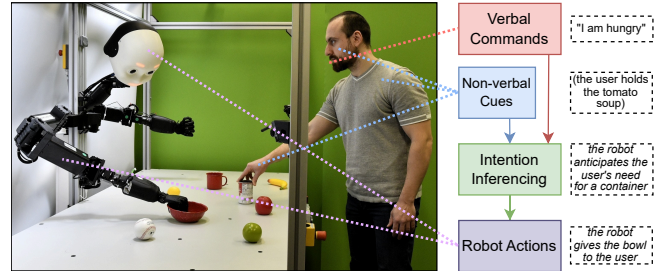


Fig. 1. An overview of our intention prediction system. After inferring both verbal and non-verbal cues using an LLM, the robot generates appropriate actions, e.g., giving the bowl to the user.

prediction in HRI. However, LLMs heavily depend on textual input (user queries and prompts) to generate contextually relevant responses. Thus, when used with robots, it remains challenging for LLMs to replace human-aware models during the interaction [5], even with the subsequent generation of multimodal LLMs, which still face difficulties processing and understanding visual information and basic visual patterns [6]. Thus, a reliable approach for equipping LLMs with awareness of human non-verbal cues is still missing.

In this work, we provide the Neuro-Inspired COLlaborator [7] (NICOL) robot (see Fig. 1) with the capacity to predict human intentions and anticipate actions by integrating several verbal and non-verbal cues with an LLM-powered intent reasoning framework. By following a two-layered processing approach, referred to as *perceptive reasoning* and *task reasoning*, we process user non-verbal cues, captured with our machine learning vision models. Then, we use a text-based LLM's inferential skills to identify current user actions relevant for the collaborative environment and anticipate user goals behind the inferred actions. We evaluate our system through a collaborative task for object categorization. Our experiments confirm the potential of LLMs as viable assets for building a competent intention prediction system. Also, we showcase the importance of non-verbal communication for an intuitive and natural collaboration with the robot. Overall, our work contributes to both research in LLMs and robotics as well as human intention prediction in human-robot collaboration, effectively bridging the gap between machine understanding and the subtleties of human communication.

II. RELATED WORK

Predicting user intention in HRI is a considerable challenge, given the immense individual human differences and the complexity of dynamic environments [3]. Also, efficient prediction of human intention requires intensive comprehen-

¹The research was supported by Horizon Europe project TERAIS, under Grant agreement number 101079338 and the DFG under the Cross-modal Learning (TRR-169) project. The authors are with the Knowledge Technology Group, Department of Informatics, University of Hamburg, Germany. Emails: {hassan.ali, philipp.allgeuer, stefan.wermter}@uni-hamburg.de

Input:
 USER: Tell me about yourself.
Output:
 SAY: Hello! I am NICOL, short for "Neuro-Inspired Collaborator". I am a humanoid robot designed and built by the Knowledge Technology group of the University of Hamburg, and I am here to collaborate with you!

Input:
 USER: Give me the red object which is also a fruit.
Output:
 SAY: I see that you're referring to the apple. Sure I can give it to you.
 ACTION: Give apple

Prompt 1: Our system is based on a grounded LLM as the NICOL robot [20].

sion of context and continuous awareness of the user [8]. Besides LLMs' proficiency in language comprehension, they have shown promising capabilities in contextual understanding [9]. They also exhibit adaptability to user input and requirements, while possessing potential for processing real-world knowledge [10]. Thus, LLMs are potential assets for tasks including real-time coordination between a human and an AI agent [11], trust-based human-aware zero-shot planning [5], belief-desire-intention driven planning [12], and human-aware planning and action anticipation [13]. LLMs have also been useful in anticipating human actions while tuned in a few-shot learning fashion [14], and employed as tools for solving ambiguity in human instructions through multiple dialogue rounds for clarification of intentions [15].

While explaining intentions with verbal cues and conversation repair is beneficial, incorporating non-verbal communication attributes to a seamless and intuitive interaction, thereby increasing human awareness in AI systems. Prior conducted work primarily utilized non-verbal cues for providing robots with human-like capabilities for apparent enhancement of the HRI experience. Some research included grounding LLMs in humanoid robots to generate movements and poses [16] and improving the embodiment of a conversational agent with a focus on engagement through features like facial expressions [17]. Other approaches focused on improving the interaction liveliness by hand gestures [18]. Work was also done on using the robot eye gaze for managing dialogue while interacting with multiple users [19]. However, there remains a notable gap in leveraging human non-verbal cues for the exact purpose of improving intention prediction. In our work, we demonstrate that integrating non-verbal cues with LLMs contributes to accurate predictions of human intentions, ultimately enriching the interaction quality.

III. METHODOLOGY

We base our system on a robustly grounded ChatGPT LLM [20], mirroring the behavior and interaction style of a physical robot, i.e., NICOL robot, rather than interacting with users as an AI language model. This is achieved by equipping the LLM with facts about the environment through the robot sensors, and linking the LLM abstract reasoning with the robot physical actions in a modular open architecture. An example response by the said LLM is in Prompt 1. The framework processes user queries and generates responses in first person, encoding robot actions without explicit programming. These LLM-generated responses are parsed into tokens recognized by NICOL's ROS-based API platform,

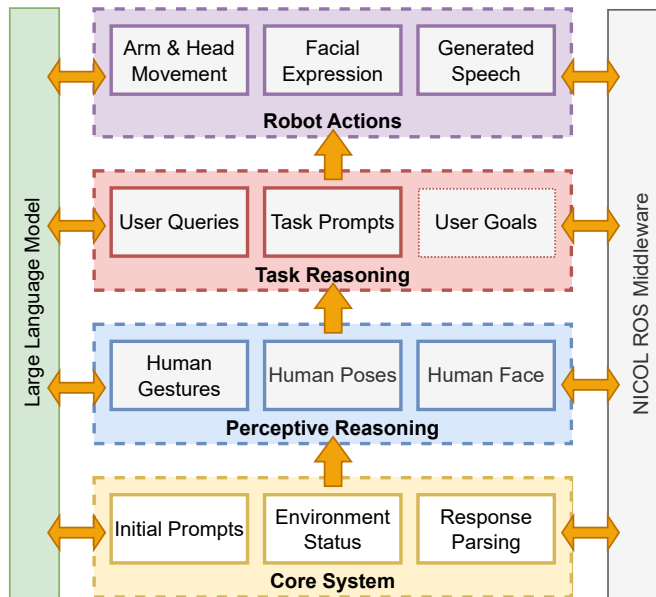


Fig. 2. Our proposed system hierarchy for intention prediction.

then mapped into concrete actions (ex: *SAY* to generate speech or *ACTION* to generate a robotic manipulation).

Our framework (see Fig. 2) utilizes the described concept [20] to implement and operate through a dual process for intention prediction: *perceptive reasoning* and *task reasoning*. Perceptive reasoning refers to the robot's ability to collect and analyze information about the surrounding environment and interacting user, thus formulating a set of "beliefs" guiding its behavior. In contrast, task reasoning involves incorporating user guidance, explicit commands, and contextual factors to collaborate towards a mutual goal. These inputs are integrated with the beliefs created with perceptive reasoning to facilitate effective decision-making and goal attainment. By seamlessly processing the information through the described layered hierarchy, our framework enhances the robot's adaptability and responsiveness towards various non-verbal user cues, while effectively interacting with objects on the table and generating useful robot actions.

A. Non-verbal Cues

Building a human-aware robotic system requires perceiving not only user speech but also user non-verbal cues [21], conveying human intentions. Therefore, it is essential to enable our system to detect and interpret human gestures and body language. To meet the real-time demands of robotic systems, we leverage MediaPipe [22] to integrate various perception components. MediaPipe offers a machine learning pipeline that facilitates a GPU integration, thanks to the recently released open-source Tasks API. Although it was originally developed with a focus on mobile devices, our experiments showcase its versatility and cross-platform suitability for a Python integration with a robot in a tabletop setup. Essentially, MediaPipe constructs a computer vision pipeline consisting of directed graphs of modular components. These graph nodes manage the data flow through data streams, configured with increasing timestamps. Each

processing node performs a specific calculation like image transformation, limb detection, limb tracking or a machine learning inferential task using frameworks like Tensorflow and Caffe, transforming the sensory input into meaningful outputs like body landmarks. We construct three different pipelines to extract various non-verbal cues:

a) *Hand Pipeline*: this pipeline is based on a single-shot Convolutional Neural Network (CNN) for identifying hand regions, and a regression-based hand tracker, which predicts landmarks within a cropped area around the detected hand. The model predicts 21 keypoints, derived from the GHUM [23] model for human shape and poses, including four for each finger and one for the wrist. We utilize these landmarks to train a recognition model for hand gestures, including pointing, holding an object, and resting. Although the detector is capable of locating hands in various distances, we distinguish between *non-proximity* and *proximity* hand gestures, happening within a defined physical area around the robot’s workstation. We use the hand 3D coordinates to identify these hand gestures before LLM prompting.

b) *Human-Pose Pipeline*: an estimation of user poses is done in a top-down way by first detecting the human in the frame with an optimized MobileNetV2. Then, poses are estimated with 33 landmarks representing the joints with a light-weight CNN, called the BlazePose [24] model. The extracted skeleton keypoints are mapped into 3D coordinates of the torso, arms, legs, and head. We utilize the generated 3D keypoints to track the user’s location in real-time relative to the robot. Thus, we enable our system to lively interact with the user based on their location in the workspace.

c) *Face Pipeline*: a face detector outputs 478 keypoints representing 3D face landmarks. The pipeline consists of a BlazeFace model based on a Single Shot Detector [25] (SSD), which identifies the main face region representing the user’s eyes, mouth, nose, converted into a face mesh model. Also, a Blendshape model predicts 52 scores representing various facial features, used to recognize the user’s facial expressions. Thus, we enable the system to identify basic facial expressions, such as smiling, and consequently the robot to react based on the given user emotional state.

B. Perceptive Reasoning

As previously mentioned, we base our intention understanding mechanism on a grounded LLM [20]. When such a system is booted, a set of initial prompts are executed, responsible for providing the LLM with information about the robot, its capabilities, surrounding environment, and purpose. The LLM is also provided with a real-time stream of information about the objects on the robot’s table using the ViLD [26] open-vocabulary object detector as well as user utterances with an ASR system based on Whisper [27]. A set of robot actions are generated like object and head manipulation based on the prompts output. We use this LLM backbone for predicting the user’s intention by collecting user verbal commands, non-verbal cues, and the environment state and converting them into texts used to prompt the LLM.

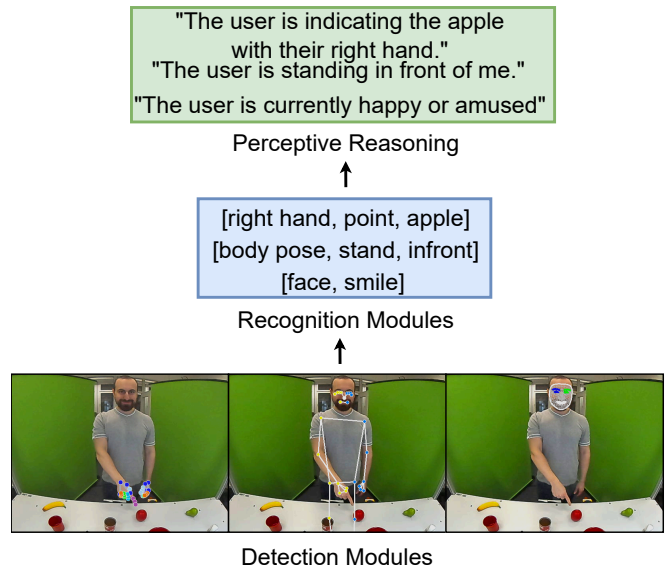


Fig. 3. Perceptive reasoning of user non-verbal cues with examples.

Essential to this reasoning step is the system’s ability to perceive human actions and gestures, thus being constantly “human aware”. For that, we develop a so-called perceptive reasoning method, in which the system interprets non-verbal cues, namely: hand gestures, body poses, and facial expressions, and combine them with the environment state to make an appropriate decision leading to a collaborative action. We follow a hierarchical approach for parsing and interpreting the cues before LLM prompting. First, the camera input passes to the various detectors (ex: hand detector), where joint keypoints are extracted and then passed into a recognition module (ex: gesture recognition). The recognition output provides basic textual tokens about the non-verbal cues, e.g., the detected limb and recognized movement. Later, these tokens pass into the LLM as internal prompts invisible to the user to be converted into meaningful textual representations of the user’s state. Fig. 3 highlights the steps with concrete examples. For this step, we consider proximity gestures only.

C. Task Reasoning

During user-robot collaborations, the system considers the user states extracted in the perceptive reasoning step while acting to achieve a mutual goal. Since we evaluate our system using object categorization, this task will be the ultimate goal steering the interaction. Tasks are described in the form of *task prompts*, which are exchangeable and expandable in a modular way. In task reasoning, the system combines the user states, user explicit commands, i.e., user queries, as well as any given task prompts in order to prompt the LLM. If an additional goal is defined by the user, then the system will use the given information to achieve it. Otherwise, the system will respond to the user based on the available information as the user intentions and goals are often vague and require multiple iterations to clarify (thus the dotted line in Fig. 2).

Prompt 2 displays an example of a defined task prompt for categorizing objects. By following the proposed hierarchical architecture, the system can infer user goals and intentions

SYSTEM: In this demo, you and the user should collaboratively split the objects on the table into two categories. The location of the object on the table represents its category. There are three different possibilities (right, centre, left). Category 1 represents all the objects on the right side of the table, and Category 2 represents all the objects on the left side of the table. Objects in the centre do not belong to neither categories yet and are still uncategorized. You will not be told the exact categories explicitly but you should try to infer that from the actions of the user. The categories will be based on known properties of the objects like shape, purpose, material, and color. As you understand better the categories that the user is trying to create, you should help the user decide in which category an object belongs.

Prompt 2: An example of a defined *task prompt* for categorizing objects.

- the user placed the can on the right side of the table and apple on the opposite -
Input:
 CUE: The user is pointing to the banana
 USER: Considering the properties and usage of this object, which of category 1 or category 2 is more fitting?
Output:
 Considering that the banana is a fruit, it would be more fitting in Category 2, which currently contains other fruits.
 ACTION: Move yellow banana to the left side of the table

Input:
 What is your inference about the user intention from the previous interactions?
Output:
 From our previous interactions, I infer that the user intended to categorize the objects on the table into two groups: containers and fruits.

Prompt 3: The robot infers the user’s goal when categorizing objects.

Input:
 CUE: The user holds the tomato soup can
 USER: I am hungry
Output
 ACTION: Give bowl to user
 SAY: Would you like me to pour the tomato soup into the bowl for you?

Prompt 4: The robot reads user intention and anticipates the next action.

in a categorization task (see Prompt 3) and during object interactions (see Prompt 4). Nevertheless, our system exhibits various additional useful features, such as establishing a joint object attention between the user and robot, utilizing the LLM’s language comprehension and generation skills to perform conversation repairs, identify object ambiguities (see Prompt 5), all while interacting naturally and intuitively with hand gestures. The robot can also identify user proximity, i.e., their location relative to the robot, thus initiating the interaction and playing both active and passive roles, while reacting socially to user facial expressions (cf. Prompt 6).

IV. EXPERIMENTS AND RESULTS

A. NICOL Robot & Environment Overview

We use NICOL [7], which is a semi-humanoid robot developed at the Knowledge Technology Group of the University of Hamburg. NICOL is an adult-sized robot fixed on a tabletop with two arms and a humanoid robotic head. Each arm has a five-fingered hand, is capable of bio-inspired motion planning [28], and can interact with adult-sized objects. The design of the NICOL robot allows collaboration with humans through precise physical interactions with the environment. Nevertheless, it also features social interaction with robot cues like facial expressions, gestures, and gaze.

B. Object Categorization Task

In our task, the user and robot collaboratively categorize objects on the table into two distinct groups. The robot assists the user in finding the fit for each “uncategorized”

Input:
 CUE: The user points to the banana
Output:
 ACTION: Look at banana
 SAY: The banana is a fruit that is usually yellow when ripe. It is a good source of potassium, vitamin C, and dietary fiber.

Input:
 USER: Give me the yellow object
Output:
 SAY: Sure, I’d be happy to help! Could you please clarify which object you’re referring to as the yellow one? Is it the lemon or the banana?

Input:
 USER: The user points to the banana
Output:
 SAY: Ah, I see! You’re referring to the banana. Sure, I can give you the banana. Here you go!
 ACTION: Give banana to user

Input:
 USER: Give me another object
Output:
 SAY: Certainly, which object would you like me to give you? Just let me know and I’ll be happy to assist you.

Input:
 CUE: The user points to the lemon
Output:
 SAY: Certainly, I can give you the lemon.
 ACTION: Give lemon to user

Prompt 5: Object disambiguation and intuitive interaction using pointing gestures.

Input:
 CUE: The user has both hands rested on the table
Output:
 ACTION: Look at user
 SAY: Hello there! How can I assist you today?

Input:
 CUE: The user is pointing to the table, The user is smiling
 USER: Hi
Output
 ACTION: Express happiness
 SAY: Hello! How can I assist you today? Would you like me to look at or point to a specific object on the table? Or perhaps you have a question for me?

Prompt 6: The robot detects user proximity and initiates an interaction accordingly.

Phase 1 (Task Understanding Check):
 USER: What are the objects in category 1?
 USER: What are the objects in category 2?
 USER: What do objects in category 1 have in common?
 USER: What do objects in category 2 have in common?
 USER: Which objects are currently uncategorized?
Phase 2 (Object Categorization Check):
the following prompt is repeated four times:
 CUE: The user is pointing to the *object_name*
 USER: Considering the properties of this object, which of category 1 or category 2 is more fitting?
Phase 3 (Intention Understanding Check):
 USER: What do you infer about the user intention from the previous interactions?
 USER: Can you explain in detail all the choices you made for categorization?

Prompt 7: Prompts used for system evaluation in an object categorization task.

object based on common properties and characteristics, like shape, color, and purpose. Unlike typical autonomous object categorization tasks [29][30][31], the robot is not given an explicit definition of the categories but should infer the goal based on user actions. The task starts as the user places one object on each side of the table, i.e., representing *Category 1* and *Category 2*. Then, the robot responds to user queries regarding where to place a specific uncategorized object. An explanatory example of the categorization task is in Fig. 4.

We evaluate the task using six objects: red apple, yellow banana, red can, yellow lemon, red bowl, and red cup. We select the objects due to their various purposes (ex: fruit

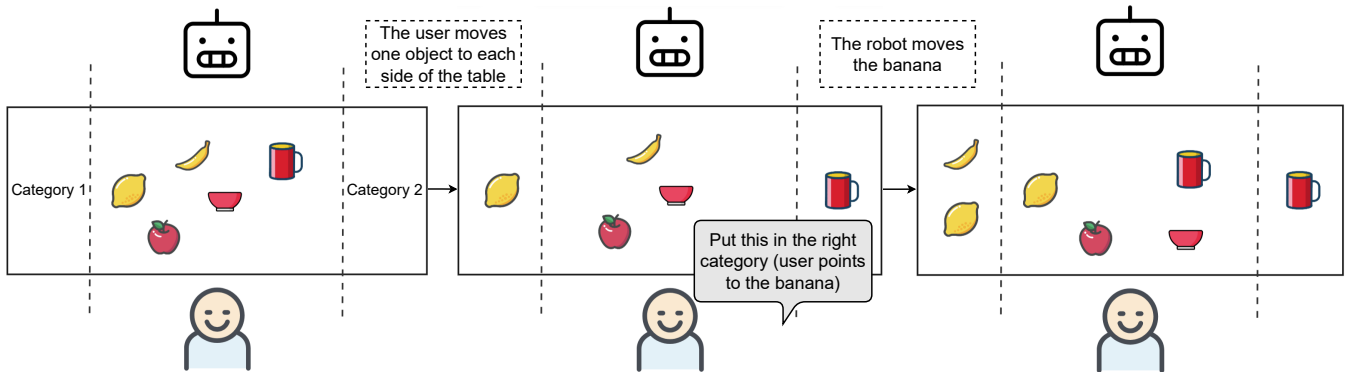


Fig. 4. A concrete workflow example of the object categorization task generated by our system.

TABLE I

RESULTS OF THE OBJECT CATEGORIZATION TASK.

	Model	Fruit container	Fruit fruit	Container fruit	Container container
Phase 1	gpt-3.5	0.99	0.68	0.91	0.88
	gpt-3.5-16k	0.62	0.74	0.89	0.81
	gpt-4	0.95	0.84	0.9	0.98
	vicuna	0.33	0.33	0.42	0.39
	mistral	0.32	0.18	0.21	0.33
Phase 2	gpt-3.5	0.82	0.52	0.8	0.61
	gpt-3.5-16k	0.7	0.73	0.98	0.78
	gpt-4	0.96	0.93	0.97	0.99
	vicuna	0.54	0.48	0.73	0.53
	mistral	0.725	0.78	0.35	0.66
Phase 3	gpt-3.5	0.71	0.37	0.71	0.7
	gpt-3.5-16k	0.75	0.53	0.89	0.63
	gpt-4	0.92	0.75	0.92	0.93
	vicuna	0.17	0.56	0.48	0.29
	mistral	0.75	0.53	0.44	0.4

and containers), colors (ex: red and yellow), and shapes (ex: round and cylindrical). For performance comparison, we start the task with every possible pair of objects as category initiators. After category initialization, the robot is prompted to categorize a randomly selected object. We conduct 150 trials in the experiment with 10 for each possible object pair, each with a reset chat history. Moreover, we evaluate the system using OpenAI and open-source LLM backbones, namely: ChatGPT gpt-3.5-turbo-0301, gpt-3.5-turbo-16k-0613, gpt-4-0613, vicuna-1.3-33b, and mistral-7b-instruct-v0.2. We follow a three-phase evaluation scheme, with a total of 11 prompts (see Prompt 7): 1) Task Understanding Check: five prompts testing the LLM’s understanding of the task and categories at hand, 2) Object Categorization Check: for each “uncategorized” object, a prompt is given in the form of a categorization request (four prompts in total), and 3) Intention Understanding Check, in which the LLM’s understanding of user intentions and goals is tested with two prompts. The system is also expected to provide logical explanations about its decisions and actions. A summary of the results is in Table I with values representing model accuracy, averaged by the number of trials and object pairs.

Our system exhibited proficiency in predicting user goals and deriving intended categories of the various fruits and containers, irrespective of the comparison order, especially gpt-4, with highly consistent accuracy. The performance of the gpt-3.5 models was comparable and dropped slightly

TABLE II

PEARSON CORRELATION OF CATEGORIZATION PHASES (WITH p -values).

Model	Phase 1–Phase 2	Phase 2–Phase 3	Phase 1–Phase 3
gpt-3.5	0.8 (0.0003)	0.6 (0.018)	0.53 (0.044)
gpt-3.5-16k	0.53 (0.04)	0.51 (0.05)	0.4 (0.14)
gpt-4	0.44 (0.1)	0.79 (0.0005)	0.63 (0.012)
vicuna	0.53 (0.043)	0.75 (0.001)	0.49 (0.062)
mistral	0.56 (0.029)	0.21 (0.44)	0.26 (0.35)

when comparing objects with similar function (fruit vs. fruit or container vs. container), indicating these models gave more importance to “usage” over other attributes. Nonetheless, gpt-4 demonstrated competence and flexibility when the categorizing criteria depended on subtle features like color or shape. All OpenAI models showed good task realisation in Phase 1 and the occasional accuracy drop is explained by the accidental mixing of the category location (swapping right and left), hinting that the LLM spatial reasoning and orientation have not yet progressed to a performant level.

The open-source models came short in Phase 1 while having near to above-average accuracy in Phases 2 and 3. Mistral showed higher competence in Phase 2 than Vicuna despite the smaller size (7B vs. 33B parameters), which we hypothesize to be a result of its own training, i.e., Mistral is fine-tuned on instructions data making it more flexible with user commands, while Vicuna is fine-tuned on conversation data. However, our results showed a positive correlation between the model’s task understanding in Phase 1 and its performance in subsequent phases (cf. Table II). This becomes clearer by observing the error rates (see Fig. 5), which are lower for “Category Reasoning” than grounding factors (category definition and uncategorized objects). Also, all LLMs showed high success rates at interpreting non-verbal cues through our perceptive reasoning. From the OpenAI models, gpt-4 reached the best performance at Phase 3, while the other OpenAI models, despite making correct decisions, were less efficient at precise decision explaining. Although ChatGPT models showed higher resilience in grounding to a new task, the results showcased that the highest error rates were related to not fully grasping some task-related concepts rather than the categorization decision itself, suggesting that the model’s reasoning aligns with the objectives of the task. As the model develops a better understanding of the task over time, its performance is expected to improve accordingly.

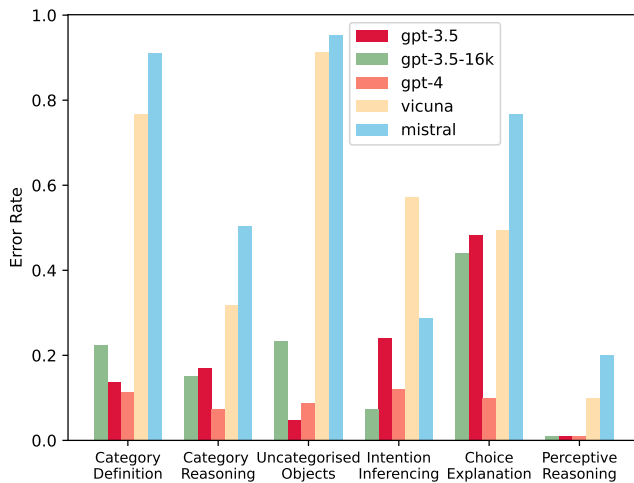


Fig. 5. Sources of system errors in the object categorization task.

V. DISCUSSION AND CONCLUSION

We explored LLMs for predicting human intention, showing their ability to infer user goals and generate discrete actions to fulfill them in a human-robot collaborative task. We proposed a hierarchical process for intention prediction with our perceptive and task reasoning modules. We illustrated the high capacity of LLMs to interpret non-verbal cues through our perceptive reasoning. Our system was also able to adapt to a given task, i.e., object categorization, in a few-shot manner, highlighting its potential for rapid task acquisition and deployment in diverse settings, thus laying the groundwork for effective intention prediction in a robotic collaborative scenario. Future research endeavors will assess the LLM scalability across a broader range of interactive scenarios to ascertain the transferability of its intention prediction skills across multiple tasks. Our future work will also examine the applicability of multimodal LLMs for inferring user gestures and poses in real-time interaction with the robot.

REFERENCES

- [1] E. Hildt, "What sort of robots do we want to interact with? reflecting on the human side of human-artificial intelligence interaction," *Frontiers in Computer Science*, vol. 3, 2021.
- [2] Y. Zhang and T. Doyle, "Integrating intention-based systems in human-robot interaction: a scoping review of sensors, algorithms, and trust," *Frontiers in Robotics and AI*, vol. 10, 2023.
- [3] A. Lubitz, L. Gutzeit, and F. Kirchner, "CoBaIR: A python library for context-based intention recognition in human-robot-interaction," in *2023 32nd IEEE RO-MAN*, 2023, pp. 2003–2009.
- [4] K. Valmeekam, S. Sreedharan, M. Marquez, A. O. Hernandez, and S. Kambhampati, "On the planning abilities of large language models," *ArXiv*, vol. abs/2302.06706, 2023.
- [5] B. Zhang and H. Soh, "Large language models as zero-shot human models for human-robot interaction," *IEEE/RSJ IROS*, pp. 7961–7968, 2023.
- [6] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal LLMs," *ArXiv*, vol. abs/2401.06209, 2024.
- [7] M. Kerzel, P. Allgeuer, E. Strahl, N. Frick, J.-G. Habekost, M. Eppe, and S. Wermter, "NICOL: A neuro-inspired collaborative semi-humanoid robot that bridges social interaction and reliable manipulation," *IEEE Access*, vol. 11, pp. 123 531–123 542, 2023.
- [8] S. Veselic, C. Zito, and D. Farina, "Human-robot interaction with robust prediction of movement intention surpasses manual control," *Frontiers in Neurorobotics*, vol. 15, 2021.

- [9] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *ArXiv*, vol. abs/2312.02003, 2024.
- [10] B. Alkhamissi, M. Li, A. Celikyilmaz, M. Diab, and M. Ghazvininejad, "A review on language models as knowledge bases," *ArXiv*, vol. abs/2204.06031, 2022.
- [11] J. Liu, C. Yu, J. Gao, Y. Xie, Q. Liao, Y. Wu, and Y. Wang, "LLM-powered hierarchical language agent for real-time Human-AI coordination," *ArXiv*, vol. abs/2312.15224, 2024.
- [12] M. Jang, Y. Yoon, J. Choi, H. Ong, and J. Kim, "A structured prompting based on belief-desire-intention model for proactive and explainable task planning," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, 2023, p. 375–377.
- [13] M. A. Graule and V. Isler, "GG-LLM: Geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning," *ArXiv*, vol. abs/2310.20034, 2023.
- [14] Q. Zhao, S. Wang, C. Zhang, C. Fu, M. Q. Do, N. Agarwal, K. Lee, and C. Sun, "AntGPT: Can large language models help long-term action anticipation from videos?" *ArXiv*, vol. abs/2307.16368, 2023.
- [15] M. J. Q. Zhang and E. Choi, "Clarify when necessary: Resolving ambiguity through interaction with LMs," *ArXiv*, vol. abs/2311.09469, 2023.
- [16] T. Yoshida, A. Masumori, and T. Ikegami, "From text to motion: Grounding GPT-4 in a humanoid robot "Alter3"," *ArXiv*, vol. abs/2312.06571, 2023.
- [17] N. Cherakara, F. Varghese, and S. S. et al., "FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions," *ArXiv*, vol. abs/2308.15214, 2023.
- [18] Y. K. Lee, Y. Jung, G. Kang, and S. Hahn, "Developing social robots with empathetic non-verbal cues using large language models," *ArXiv*, vol. abs/2308.16529, 2023.
- [19] G. J. Serfaty, V. O. Barnard, and J. P. Salisbury, "Generative facial expressions and eye gaze behavior from prompts for multi-human-robot interaction," in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, NY, USA, 2023.
- [20] P. Allgeuer, H. Ali, and S. Wermter, "When robots get chatty: Grounding multimodal human-robot conversation and collaboration," 2024, submitted to ICANN 2024.
- [21] H. Ali, D. Jirak, and S. Wermter, "Snapture—a novel neural architecture for combined static and dynamic hand gesture recognition," *Cognitive Computation*, vol. 15, no. 6, pp. 2014–2033, Nov 2023.
- [22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Ubowaja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, and W.-T. C. et al., "MediaPipe: A framework for building perception pipelines," *ArXiv*, vol. abs/1906.08172, 2019.
- [23] H. Xu, E. G. Bazavan, A. Zafir, B. Freeman, R. Sukthankar, and C. Sminchisescu, "GHUM & GHUML: Generative 3D human shape and articulated pose models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*, 2020, pp. 6184–6193.
- [24] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," *ArXiv*, vol. abs/2006.10204, 2020.
- [25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37.
- [26] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *ArXiv*, vol. abs/2104.13921, 2022.
- [27] A. Radford, J. W. Kim, T. Xu, and G. B. et al., "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.
- [28] J.-G. Habekost, E. Strahl, P. Allgeuer, M. Kerzel, and S. Wermter, "Cycleik: Neuro-inspired inverse kinematics," in *ICANN 2023*, Sep 2023, p. 457–470.
- [29] M. Koskinopoulou, F. Raptopoulos, and G. P. et al., "Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste," *IEEE Robotics & Automation Magazine*, vol. 28, no. 2, pp. 50–60, 2021.
- [30] Z. Pan and K. Hauser, "Decision making in joint push-grasp action space for large-scale object sorting," in *IEEE ICRA*, 2021, pp. 6199–6205.
- [31] S. Ramadurai and H. Jeong, "Effect of human involvement on work performance and fluency in human-robot collaboration for recycling," in *ACM/IEEE International Conference on HRI*, 2022, pp. 1007–1011.