Human Robot Collaboration with Few-Shot LLM Robot Models

Maitrey Gramopadhye
University of North Carolina
Chapel Hill, North Carolina, USA
maitrey@cs.unc.edu

Daniel Szafir
University of North Carolina
Chapel Hill, North Carolina, USA
daniel.szafir@cs.unc.edu

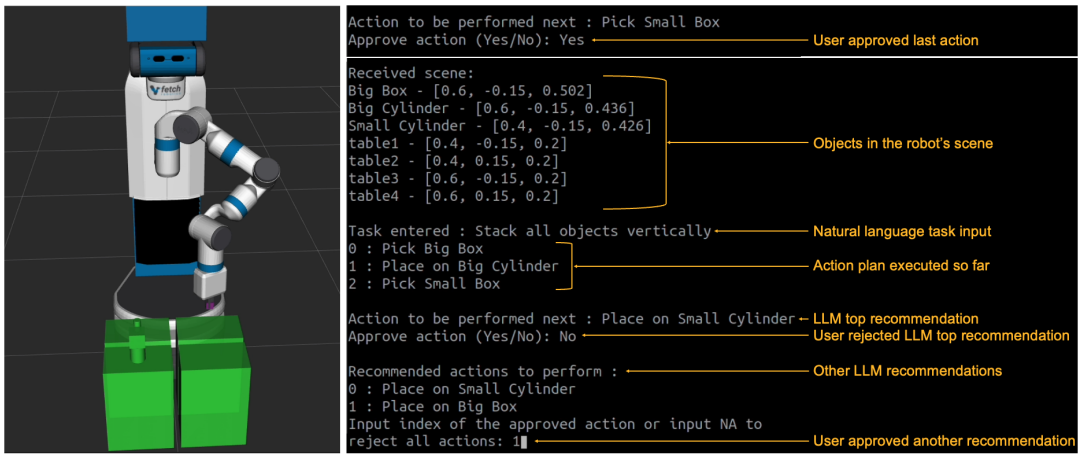

Figure 1: The robot and user interface used for our experiments. (Left) The simulated Fetch robot rendered in Rviz was used to perform *pick* and *place* operations on tabletop objects. (Right) The terminal interface enabled participants to relay natural language tasks to the Large Language Model, and examine and approve the LLMs output actions for execution on the robot.

ABSTRACT

Large Language Models have had a notable surge in popularity in task planning for embodied robots. However, most approaches have the robot operating in isolation with minimal collaboration with humans. In this paper, we design a system that enables people to interact with an intelligent robot. We conduct a human subjects study to gain insights into the participants' mental model, and whether the comprehensive abilities of LLMs encourage the users to adopt a collaborative role when working with the robot for long-horizon tasks.

1 INTRODUCTION

Large Language Models (LLMs) and Vision Language Models (VLMs) have demonstrated noteworthy reasoning capabilities, with research showing their generalization ability across areas of robotics including Task Planning [3, 4, 7, 15, 19], Policy learning [14], Robot locomotion [11, 16–18], Generating simulations for robot learning [23] and user interface for robots [2, 22]. Particularly, internet-scale models, fine-tuned using robot trajectory data, have demonstrated promising results in translating high-level tasks into primitive actions executable by an embodied robot [1, 9, 13]. While these models generate sensible outputs aligned with the robot's capabilities and environment, they frequently lead to independent robot actions, lacking interaction with people and neglecting human feedback and preferences.

Approaching the problem through the lens of Human-Robot Interaction, we propose a system to facilitate human interaction with an intelligent robot. Our system features a simulated Fetch robot [24] proficient in executing the primitive actions *pick* and *place* that is integrated with an LLM approach [9], which, when provided with a high-level task and information about the robot's environment, produces an action plan encompassing the primitive actions. Users input their high-level tasks as natural language and, for each of the actions output by the LLM, are given the option to either approve it for execution on the robot or select alternative actions, fostering a collaborative interaction.

We have evaluated our system via a preliminary user study, where participants supervised the robot's execution of various object rearrangement tasks. We specifically studied tasks that required planning over multiple steps to complete, also known as long-horizon tasks. In reviewing existing literature on Human-Robot collaboration and mental model alignment for long-horizon tasks, we observed that several studies characterized the robot model using a fixed architecture, assuming compatibility with the user's mental model [5, 6, 10, 12, 20, 21, 25]. Given the recent popularity of LLMs as robot models, our user study delves into the cognitive processes and mental models of users when engaging with an embodied robot driven by a LLM.

2 SYSTEM IMPLEMENTATION

Our fully implemented system currently supports user collaboration with a Fetch Robot [24]. We chose Fetch because of its capabilities as a general-purpose mobile manipulator, allowing it to work with users on a variety of tabletop manipulation tasks. Our current system uses a simulated version of Fetch, rendered in RViz with the same motions and controls as the real robot, to speed up experiments and ensure the safety of participants. We connected

the simulated robot to Moveit! Task Constructor (MTC) [8] to implement the primitive actions and access the robot's environment information. We implemented functions to pick up any specified object and place any picked object on a specified object. The object information relayed by MTC consists of a comprehensive list of the object names around the robot, complete with their respective 3D locations. We hosted the simulated robot and MTC as ROS nodes.

Our system builds on the LLM approach outlined by Gramopadhye et al., 2023 [9] to convert a given task and the robot's environment information to a series of primitive actions. Upon receiving the inputs, an LLM scans a pre-collected dataset of (task, action plan, robot environment) tuples and forms a suitable natural language prompt. We then iteratively query the LLM for a step-by-step action plan for the robot to complete the specified task. For each step of the action plan, we sample several recommendations from the LLM, ranked by the LLM's preference, for the user to choose from (for more details on the LLM method, see [9]).

We designed a terminal-based user interface to display important information from MTC and the LLM and cue users for input. The interface was also run as a ROS node.

3 EXPERIMENT DESIGN

We performed a pilot in-person, between-participants experiment studying how people interacted with a LLM-based robot to complete long-horizon object rearrangement tasks. We recruited 3 participants from a university campus. We varied the initial size of the dataset available for our LLM method across participants (either 0, 5 or 20 samples).

3.1 Experiment Setup

The robot had 4 tables in front of it (Figure 2 (Left)), with some or all of them having one of the objects shown in Figure 2 (Right). We reset the robot's environment between the experiment stages (discussed in §3.2).

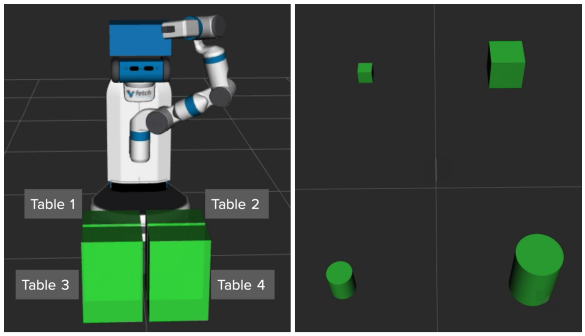

Figure 2: Experimental setup details. (Left) The simulated robot always has 4 tables in front of it. (Right) The 4 types of objects possible in the robot's environment are - Big Cube, Small Cube, Big Cylinder and Small Cylinder.

3.2 Stages

Our experiment consists of three stages with objectives of varying complexities. We communicated the objective to participants verbally, using the same phrases for each participant. However, participants were free to use whatever language they chose when interacting with the system.

3.2.1 Stage 1: "Swap the objects with the tables". In the first stage, two of the tables nearest to the robot had objects randomly placed on them and the participants were instructed to get the robot to swap the objects' locations.

3.2.2 Stage 2: "Stack similar objects together with smaller objects on the top". In this stage, all the tables in front of the robot had different objects placed on them in random order and participants were instructed to get the robot to pick-up the smaller objects (i.e., Small Cube and Small Cylinder) and place them on the larger objects of the same kind (i.e., Large Cube and Large Cylinder respectively).

3.2.3 Stage 3: "Make a tower of all objects". This stage was initialized similar to *Stage 2* and the participants were instructed to get the robot to stack all objects on top of each other in any order.

3.3 Hypotheses

This study allowed us to analyse the thought process of participants interacting with an intelligent robot, capable of understanding long-horizon tasks. We hypothesized that:

- **H1: There will be a common pattern to how participants input the task into the interface.**
- **H2: The participants will view the robot as a collaborator.**
- **H3: The size of the dataset available to the LLM will be proportional to the robot's task planning performance.**

3.4 Procedure

Our experiment consisted of four phases: (1) Introduction and procedure explanation, (2) Experiment stages, (3) Exit Questionnaire, (4) Exit Interview. After receiving an overview of the user interface and the experimental procedure, participants were given a description of their role as a collaborator. They were also shown a video of an example interaction with the system with a dummy task. Each participant was asked to complete all three stages in order each time. For each stage, our user interface cued the participant for a natural language task input (high-level instruction). Simultaneously, we queried MTC for the robot's environment. We then prompted the LLM for the step-by-step action plan and presented the output of the LLM to the participant and queried them for their choice of action to be executed by the robot. We first presented the LLMs top recommended action. If the participant rejected the action, they were subsequently presented, in order, with all the recommendations of the LLM and then the predefined set of all possible actions that the robot could take.

We relayed the user-selected action to MTC for execution, concurrently displaying the simulated robot carrying out the action in real time. After the action's completion, MTC transmitted the updated object information along with the details of the executed action to the LLM. The LLM was then prompted for the subsequent action, and this iterative process continued until the predefined instruction was completed. Following completion, the instruction, the generated action plan, and the object information were augmented to the dataset for the LLM. Simultaneously, the participant was cued for a new instruction to continue the interactive process.

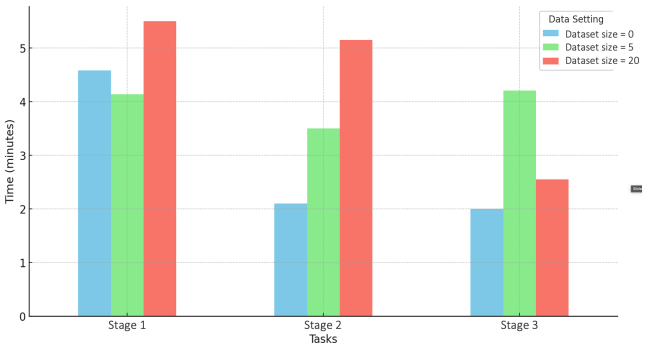

Figure 3: Comparative analysis of the time taken to complete each stage under three different initial dataset sizes of 0, 5 and 20.

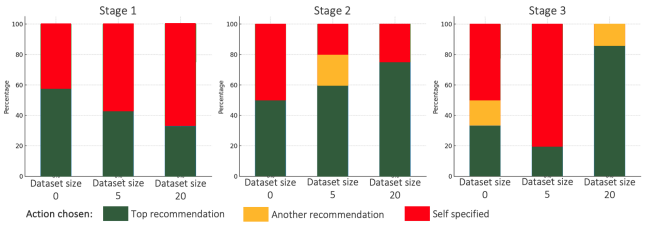

Figure 4: Distribution of participant's choice of action in terms of 'Top recommendation' (i.e. the action most recommended by the LLM), 'Another recommendation' (i.e. an action recommended by the LLM, but not the top recommendation), and 'Self specified' (i.e. action independent of the LLM recommendation) across three dataset sizes for each stage. The percentages reflect participant approval rates of the robot's action selection.

For each stage, we recorded the time elapsed, the number of instructions used to complete the stage, and the actions chosen by the participant. After the experiment, participants filled out a survey with 7-point Likert-style questions and participated in a semi-structured interview regarding their experience.

4 OBSERVATIONS

While our experiment currently contains too few participants to provide definitive conclusions, below we review our current data. We analysed the time elapsed, the number of instructions used to complete the stages and the actions chosen by the participant to determine if the size of dataset available to the LLM affects its performance.

4.0.1 Time taken: Figure 3 shows the time taken to complete the stages under varying data settings. The findings suggest a potentially complex relationship between the quantity of data provided to the LLM, the input instruction, and the time efficiency.

4.0.2 Number of instructions per stage: Participants were free to use multiple input instructions to complete a stage. However, so far all participants have used just one initial instruction for every stage. Our survey data suggests that when interacting with the robot, participants' goal was to complete the task (M = 6.33, SD = 0.577) rather than to teach the robot (M = 1, SD = 0).

4.0.3 Action chosen by participants: We observed varied response patterns from the participants in relation to the dataset size (Figure 4). We see a possible trend indicating a general increase in the tendency to approve the robot's proposed actions with increasing dataset size and stage complexity.

Our current observations in terms of our hypotheses are:

H1: We observed varied responses from participants for the input instruction. The only pattern we could observe was the participants' choice to use only one input instruction for each stage. Therefore, our current data does not support **H1**. In their interviews, participants mentioned different strategies for entering the instructions. Some participants mentioned specifying the actions to follow - "I did sub tasks such as move objects to complete the whole task, but then realized that it needed to broken down even further", whereas some entered a description of the result after task completion - "I mostly tried to communicate what the end result should look like rather than what the robot should do". Participants also mentioned that their opinion of the robot and its abilities changed over time (M = 4.66, SD = 2.309) and that they would employ a different strategy if they repeated the experiment (M = 5.66, SD = 0.577).

H2: Participants' responses were ambiguous to whether they considered the robot to be a collaborator (M = 4, SD = 1.732) or a tool (M = 5, SD = 1.732), leaning towards the latter. They were positive that giving the robot instructions to perform is a good way to collaborate with a robot (M = 6.33, SD = 1.154), but felt that the robot needed them (M = 5, SD = 1.732) and didn't think that the robot unnecessarily asked for their approval (M = 1, SD = 0). In the interviews, participants also noted that the object rearrangement tasks in the experiment were very simple. Developing more complicated tasks, that would require longer action plans to complete, might give participants more opportunities to interact with the robot and view it as a collaborator.

H3: From our tracked metrics, especially the actions chosen by participants (§4.0.3), we are finding some preliminary support for **H3** that more data might improve the LLMs task planning for the latter stages, although we must collect additional data and perform appropriate inferential analysis to fully test this hypothesis.

5 CONCLUSION

Our system and user study allowed us to take steps in understanding people's thought process when collaborating with an LLM backed robot on long-horizon tasks. From our pilot experiment, we observed a lot of variability in people's interaction with the robot. We found some preliminary support for **H3**, but need future work to dive deeper into each hypothesis.

For our future work, we plan to redesign the experiment to include more complicated real-world collaborative tasks such as assisting a chemist with a chemistry experiment or cooking, and collect additional data with more participants.

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[2] Jake Brawer, Kayleigh Bishop, Bradley Hayes, and Alessandro Roncone. 2023. Towards A Natural Language Interface for Flexible Multi-Agent Task Assignment. arXiv:2311.00153 [cs.RO]

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817* (2022).

[5] Connor Brooks and Daniel Szafir. 2019. Building second-order mental models for human-robot interaction. *arXiv preprint arXiv:1909.06508* (2019).

[6] Yujiao Cheng, Liting Sun, and Masayoshi Tomizuka. 2021. Human-Aware Robot Task Planning Based on a Hierarchical Task Model. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1136–1143. https://doi.org/10.1109/LRA.2021.3056370

[7] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).

[8] Michael Görner, Robert Haschke, Helge Ritter, and Jianwei Zhang. 2019. Moveit! task constructor for task-level motion planning. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 190–196.

[9] Maitrey Gramopadhye and Daniel Szafir. 2023. Generating Executable Action Plans with Environmentally-Aware Language Models. arXiv:2210.04964 [cs.RO]

[10] Bradley Hayes and Brian Scassellati. 2016. Autonomously constructing hierarchical task networks for planning and human-robot collaboration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 5469–5476. https://doi.org/10.1109/ICRA.2016.7487760

[11] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. 2023. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10608–10615.

[12] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43 (2019), 309–326.

[13] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*. PMLR, 9118–9147.

[14] Zhihan Liu, Hao Hu, Shenao Zhang, Hongyi Guo, Shuqi Ke, Boyi Liu, and Zhaoran Wang. 2023. Reason for Future, Act for Now: A Principled Framework for Autonomous LLM Agents with Provable Sample Efficiency. *arXiv preprint arXiv:2309.17382* (2023).

[15] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021* (2023).

[16] Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. 2023. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*. PMLR, 2683–2699.

[17] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*. PMLR, 492–504.

[18] Chak Lam Shek, Xiyang Wu, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. 2023. LANCAR: Leveraging Language for Context-Aware Robot Locomotion in Unstructured Environments. arXiv:2310.00481 [cs.RO]

[19] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11523–11530.

[20] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports* 1 (2020), 259–267.

[21] Takayoshi Takayanagi, Yusuke Kurose, and Tatsuya Harada. 2019. Hierarchical Task Planning from Object Goal State for Human-Assist Robot. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. 1359–1366. https://doi.org/10.1109/COASE.2019.8843257

[22] Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. 2024. Large Language Models for Multi-Modal Human-Robot Interaction. *arXiv preprint arXiv:2401.15174* (2024).

[23] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2024. GenSim: Generating Robotic Simulation Tasks via Large Language Models. arXiv:2310.01361 [cs.LG]

[24] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. 2016. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*. 1–6.

[25] Yang You, Vincent Thomas, Francis Colas, Rachid Alami, and Olivier Buffet. 2023. Robust Robot Planning for Human-Robot Collaboration. *arXiv preprint arXiv:2302.13916* (2023).