Understanding Large-Language Model (LLM)-powered Human-Robot Interaction

Callie Y. Kim*
Department of Computer Sciences
University of Wisconsin–Madison
Madison, Wisconsin, USA
cykim6@cs.wisc.edu

Christine P Lee*
Department of Computer Sciences
University of Wisconsin–Madison
Madison, Wisconsin, USA
cplee5@cs.wisc.edu

Bilge Mutlu
Department of Computer Sciences
University of Wisconsin–Madison
Madison, Wisconsin, USA
bilge@cs.wisc.edu

ABSTRACT

Large-language models (LLMs) hold significant promise in improving human-robot interaction, offering advanced conversational skills and versatility in managing diverse, open-ended user requests in various tasks and domains. Despite the potential to transform human-robot interaction, very little is known about the distinctive design requirements for utilizing LLMs in robots, which may differ from text and voice interaction and vary by task and context. To better understand these requirements, we conducted a user study (n = 32) comparing an LLM-powered social robot against text- and voice-based agents, analyzing task-based requirements in conversational tasks, including *choose*, *generate*, *execute*, and *negotiate*. Our findings show that LLM-powered robots elevate expectations for sophisticated non-verbal cues and excel in connection-building and deliberation, but fall short in logical communication and may induce anxiety. We provide design implications both for robots integrating LLMs and for fine-tuning LLMs for use with robots.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Natural language processing*; • **Computer systems organization** → *Robotics*.

KEYWORDS

Social robots; large language models; human-robot interaction

ACM Reference Format:
Callie Y. Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634966

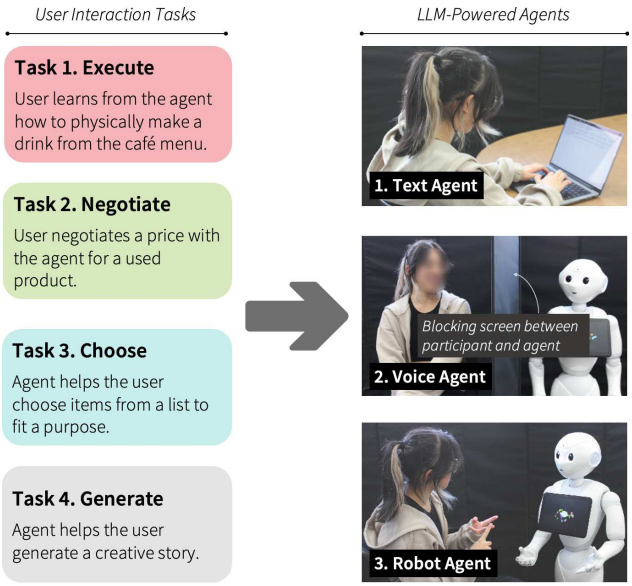

Figure 1: We investigate people's perceptions of and preferences toward LLM-powered robots. We conducted a user study that compared an LLM-powered social robot against text-based and voice-based agents. *Left:* Users participated in one of four tasks: choose, generate, execute, and negotiate. *Right:* The user engages with (1) the text-based agent by entering and receiving text-based prompts, (2) the voice-based agent through spoken prompts (achieved by the robot's voice with the robot concealed behind a black screen, out of the user's view), and (3) the LLM-powered social robot via spoken prompts, in a counterbalanced order.

1 INTRODUCTION

Across a wide range of day-to-day activities, robots are envisioned to possess social and communication skills that allow them to engage seamlessly and naturally with users [9, 34]. Past research on robots has focused on developing these skills, including conversational speech [26, 31], gestures [12, 24, 60], gaze [39, 47, 49], and appearance [20, 32, 40] to facilitate effective, continuous, and dependable interactions with users. The recent emergence of large-language models (LLMs) provides a novel opportunity for robots to augment their social and communicative abilities [70]. As these models enable lifelike conversations, contextual adaptation, and consistent interaction [10, 68], robots can leverage these capabilities to improve their communicative proficiency to effectively address diverse user requests across a range of tasks and application domains. Despite the immense potential of LLM-equipped robots to transform human-robot interaction, a gap exists in the knowledge regarding the unique design requirements for robots that harness

*Both authors contributed equally to this research.

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03…$15.00
https://doi.org/10.1145/3610977.3634966

LLMs for conversational and communicative skills, as well as which tasks most benefit from utilizing these capabilities.

Robots are known to have a unique effect on user experience and perceptions compared to other forms of embodiment, including text, voice, or virtual agents [16, 35, 41, 46, 53, 54]. Specifically, the presence of a robot triggers different cognitive activities, behaviors, or actions of a user and elicits different responses such as increased enjoyment, perceived social competence, and trust towards the robot [3, 16, 28, 42, 55, 71]. Therefore, it is conceivable that when users engage with robots powered by LLMs, the embodiment of the robot can shape distinct user expectations and perceptions of the sophisticated conversational system, which might have implications for how LLMs need to be specifically designed for human-robot interaction or integrated into a robot system.

The growing interest in integrating LLMs with robots necessitates a need to understand the unique design requirements of LLMs that are expected to work with robots, including design needs tailored to the tasks and contexts in which LLM-powered robots operate. Previous design requirements for robots have been gained through exploring user perceptions regarding various task attributes and robot roles [16, 48, 72]. This exploration can similarly uncover design opportunities and optimal tasks for LLM-powered robots, shaping future guidelines for robot design and LLM development. To understand the design requirements for utilizing LLMs for robots and identify tasks suitable for integrating LLM-powered robot agents, we formulate the following three research questions to guide this investigation: (1) how do people perceive robots using LLMs; (2) how do people's perceptions of robots using LLMs vary across different task settings; and (3) what task contexts benefit from the embodiment of a robot when people interact with LLMs?

To address our research questions, we conducted a user study with 32 participants that compared different agent types—*text*, *voice*, and *robot*—to better understand people's perceptions of LLM-powered robots compared to other forms of embodiment through which people interact with LLMs. Additionally, we designed four conversational tasks–*execute*, *generate*, *negotiate*, and *choose*, based on the "task circumplex" by McGrath [44]—to assess which tasks can benefit from LLM-powered robots. Our findings show that LLM-powered robots elicit new expectations for sophisticated non-verbal cues, and are preferred in tasks involving connection-building and deliberation between the user and the robot. Conversely, LLM-powered robots are less preferred when the LLM's rich social capabilities result in verbose responses, logical and communication errors, or induce anxiety during task interactions. Finally, we present design recommendations for LLM-powered robots to enhance future HRI. We make the following contributions:

(1) Compare LLM-powered agents (*i.e.,* text-based, voice-based, and social robot) to uncover unique design requirements for LLM-powered robots;
(2) Evaluate LLM effectiveness across tasks (*i.e.,* generate, choose, negotiate, execute) to identify optimal interaction contexts with robot embodiment;
(3) Present empirical evidence on user perceptions and preferences for LLM-powered robots in diverse task settings;
(4) Provide design implications for developing LLM-powered robots and LLMs to improve future human-robot interaction.

2 RELATED WORK

Embodiment. Embodiment plays a pivotal role in shaping how humans perceive and engage with robots. We adopt the definition of embodiment "structural coupling" from Ziemke [76] such that a system is embodied if mutual perturbative channels exist. We focus on physically embodied robots that can leverage rich channels of communication such as gesture, posture, gaze, facial expressions, proxemics, and social touch [16]. Prior research shows that interactions with physically embodied robots lead to higher user engagement, enjoyment, trust, and empathy compared to text, voice-based, or virtual agents [4, 6, 63, 67]. Additionally, embodiment influences user behavior, affecting interaction duration and distance [45, 59]. Several studies have explored how physical embodiment affects task performance and impression by comparing physically embodied robots to virtual agents [21, 46, 62, 72]. These studies indicate that user preferences for embodied agents are influenced not only by embodiment but also by the specific task context.

LLM in Robotics. Robots function as the vital bridge connecting the tangible real world and LLMs. This connection enables LLM to infer knowledge from the physical environment through data collected by sensors. Simultaneously, LLMs empower the robot with the capability to comprehend semantic meanings and engage in flexible dialogue interactions. Thus, LLMs with robots find their primary applications in task planning [1, 17, 66] or human-robot collaboration [30, 75]. For instance, Ye et al. [75] investigated the implications of LLM-powered robots when users controlled the robot through text for assembly tasks in virtual reality.

Researchers have also explored the effectiveness of LLMs for conversational robots in specific tasks. Cherakara et al. [11] designed a system in which the robot displays appropriate facial expressions when conveying information about the National Robotarium. Irfan et al. [25] utilized LLMs to create a personalized companion robot and examined the challenges associated with open-domain dialogue when interacting with older adults. Khoo et al. [29] applied LLMs to a social robot to enhance the well-being of older adults by generating empathetic responses. Yamazaki et al. [74] constructed a scenario-based dialogue system for a robot and demonstrated the effectiveness of LLMs while establishing trust with users. While prior research has primarily concentrated on evaluating the efficacy of LLM-powered robots in specific tasks, we aim to explore a wider array of tasks and contexts where LLM-powered robots can offer advantages and comprehend the unique design requirements to effectively incorporate LLMs with robots across diverse task settings.

3 METHOD

3.1 Embodiment Design

To understand people's perceptions of robots when powered by LLMs, we compare a social robot agent against two other agents–a text-based agent and a voice-based agent. All three agents were equipped with GPT-3.5, OpenAI's text-davinci-003 model [10] without fine-tuning. The model parameters were set to temperature = 0.7 with max tokens = 2048. Pre-prompts were used to outline the four tasks, with parameters identical to those used by Billing et al. [7] in Pepperchat.

Figure 2: *Interaction Examples per Each Task* — Participants were assigned to one task among the four (*i.e.,* execute, negotiate, choose, and generate) and engaged with all three types of agents (*i.e.,* text, voice, and robot.) *Top left to clockwise:* shows interaction examples of the four tasks.

3.1.1 Text Agent. Resembling a chatbot, the users interacted with the text agent through text input and output. The user sent and received prompts via the GPT model with OpenAI API.

3.1.2 Voice Agent. Simulating a voice assistant, the voice agent communicated exclusively through voice commands. For the voice agent, the participant and the robot were separated by a screen such that the participant only interacted with the agent through voice. It utilized the robot's module, "ALAudioDevice [57]" to capture the user's speech. The audio recording is then sent to Google Cloud service [19] for speech-to-text analysis, then forwarded to the GPT model via OpenAI API. The GPT model generates a response, which is converted into a speech using the robot Pepper's [58] module, "ALAnimatedSpeech [56]." The same robot was used for both the voice and robot agent instead of a smart speaker to avoid favoring one specific technology over another within the broad space of voice-based agents (*i.e.,* smart speakers, smart displays, and virtual assistants) and to ensure consistent voice interactions across both voice and robot agent conditions.

3.1.3 Robot Agent. The social robot, Pepper, was used to engage with users through animated gestures, text-to-speech, and face recognition. For successful communication between the participant and the LLM-powered robot, we employed Pepperchat [7], which utilizes Google Cloud speech-to-text functionality for speech-based dialogue, contributing to a seamless and responsive communication experience. We chose a minimalist design for the robotic agent, emphasizing its basic embodiment to highlight high-level differences among text, voice, and robot embodiments, rather than fully utilizing non-verbal cues. Thus, we chose to accept an out-of-the-box implementation of each agent, rather than each agent having specific design features (*e.g.,* visual cues for the voice agent.)

3.2 Task Design

To understand the design requirements for LLM-powered robots across various task settings, we designed different tasks based on the Group Task Circumplex Model proposed by McGrath [44]. The circumplex model is structured around two dimensions, ranging from conflict-based to cooperative, and conceptual to behavioral. The circumplex model classifies group tasks into four categories: (1) generate: tasks that involve generating ideas or plans; (2) choose: tasks that involve choosing a solution or plan from a set of alternatives where the correct or agreed-upon answer exists; (3) negotiate: tasks that involve resolving conflict of viewpoints, interests, and motives; and (4) execute: tasks that involve executing a plan or performance. This framework offers a structured approach to comprehend the nature of the tasks that groups undertake. Figure 2 shows examples of task interactions. Below we discuss the specific tasks designed for our study.

3.2.1 Generation Task. In the generation task, the agent and the participant collaboratively create an imaginary story. Participants were asked to follow a general guideline to introduce characters, features of the characters, and the setting for story development. To create the foundation and actual story, the participant and agent took turns each adding a sentence. To construct a comprehensive story, the participants were told to ideally incorporate obstacles, solutions to address the obstacles, a climax in the story, and a plot.

3.2.2 Choosing Task. In the choosing task, the agent assisted the participants in selecting a subset of items from a collection of items. There was a different theme for the collection of items for each task, including a ski, beach, and camping trip. Participants were told to select items that focused on practicality over leisure. The item criteria were based on those commonly featured as essential on various travel websites. Participants engaged in discussion with the agent to finalize their item list.

3.2.3 Execution Task. In the execution task, the agent acted as an instructor and the participant acted as a student. The agent's role was to teach the participant how to prepare a beverage in a cafe setting. Only the agent knew which drink to make and participants were asked to follow the instructions. Participants were told to ask the agent if they had any confusion or questions.

3.2.4 Negotiate Task. In the negotiation task, the agent acted as a seller of second-hand items and the participant acted as the potential buyer. The agent's goal was to sell the item as expensive as possible and the participant's goal was to buy the item as cheap as possible. The agent was not aware of how much money the participant held. To control the task settings and provide consistency, an absolute minimum price line was set for the item.

4 USER STUDY

4.1 Study Design

The study followed a mixed-factorial design with scenario tasks as the between-subjects factor and the agent embodiment as the within-subjects factor. Participants were randomly assigned to one of four tasks (*i.e.,* generate, choose, execute, and negotiate) and then engaged with the three different agents (*i.e.,* text agent, voice agent, and robot) in counterbalanced order. At the beginning of the study, participants were shown interaction examples with the LLM-powered agents that involved disagreeing with suggestions, asking follow-up questions, and tracking task progress. Additionally, the task given per agent differed slightly in topic to avoid the learning effect (*e.g.,* a camping, beach, and ski trip). Prompts for the tasks

can be found in the supplementary materials[1]. After interaction with each agent, participants completed questionnaires and a semi-structured interview about their experience. All sessions were held in person, audio and video recorded through Zoom [77] on a laptop.

4.2 Measures

4.2.1 Subjective Measures. To measure participants' perception of the agents, we used a modified version of the Godspeed questionnaire [5], which includes a series of semantic scales for measuring the robot's animacy (Cronbach's $\alpha = 0.91$), anthropomorphism (Cronbach's $\alpha = 0.89$), likeability (Cronbach's $\alpha = 0.93$), perceived intelligence (Cronbach's $\alpha = 0.91$), and perceived safety (Cronbach's $\alpha = 0.72$) on a five-point rating scale. We modified the questions such that the items asked about their perceptions of "agents" instead of "robots." Our analysis of the item reliability of the perceived safety subscale found a Cronbach's α of -0.27, due to a miscoded item in the subscale.

Godspeed items are written in a consistent way, such that in each group high values of a variable indicate a similar direction. Specifically, the miscoded variable, *still* (anchored at 1) to *surprised* (anchored at 5) appeared to differ in direction from other items: *anxious* (anchored at 1) to *relaxed* (anchored at 5) and *agitated* (anchored at 1) to *calm* (anchored at 5). After re-coding the *still-surprised* item by flipping the scale so the semantic meaning of the item would be consistent with others, we calculated Cronbach's α of 0.72. This miscoding of the *still-surprised* item and correction by reverse-coding have been reported by prior work that used the Godspeed questionnaire [*e.g.,* 2, 8, 61]. Additionally, upon closer inspection of the items of this subscale, which included anxious-relaxed, agitated-calm, and surprised-still (after reversion), we determined these items to be a poor fit to the overall construct of "perceived safety" and decided to exclude it from our analysis.

In addition to the Godspeed questionnaire, we measured participants' satisfaction (Cronbach's $\alpha = 0.96$) with the interaction on a seven-point rating scale (1 = strongly disagree; 7 = strongly agree) using the satisfaction subscales from the Usefulness, Satisfaction, and Ease of Use (USE) Questionnaire proposed by Lund [38]. The overall Cronbach's α value for the Godspeed attributes and satisfaction was 0.97.

4.2.2 Behavior Measures. To observe and understand participant behaviors, we collected measures of the total number of input tokens derived from the participants' prompts. This approach involves counting discrete units that the OpenAI API divides from the user's input to process the prompt. This metric enables assessment of the length of dialogue input provided by the user within the conversation during the task.

4.2.3 Performance Measures. To understand the quality of the interaction, we measured the number of failures that occurred during the interaction. We considered two categories of failures: (1) *technical errors*, such as interruptions by the agent, and inaccurate transcriptions from Automatic Speech Recognition (ASR); and (2) *hallucinations*, where the response from the LLM is nonsensical or unfaithful to the provided source input [27].

[1]The supplementary materials can be found at https://osf.io/exjrd/?view_only=88c0b1ff4b2b4f969928a614c9fa8fff

4.3 Participants

We recruited 32 participants (10 male, 20 female, 1 gender-queer, 1 non-binary) through a university mailing list between the ages of 18 and 59 ($M = 27.47, STD = 10.30$) where 69% were White, 28% were Asian, and 3% preferred not to answer. Participants were required to be in the United States, fluent in English, and at least 18 years old. All participants agreed to participate in our study via our institution's IRB-approved consent form. The study lasted for approximately 60 minutes and participants received $15 per hour for compensation upon study completion.

4.4 Analysis

Factorial repeated-measures analysis of variance (ANOVA) was used to determine whether the task and agent embodiment had a significant effect on all measures. If the ANOVA test showed significant effects, we tested our data for pairwise differences using Tukey honest significance test (HSD), which controls for Type I error considering all possible comparisons. The qualitative data was analyzed using Thematic Analysis (TA), following the guidelines developed by Clarke and Braun [13] and McDonald et al. [43]. The first authors became acquainted with the data by conducting the studies and initially creating a codebook [15]. Through ongoing team discussions, codes were grouped into categories and refined until a consensus was reached. These categories were then further organized and reiterated to extract themes that emerged from our study data. Once all potential themes were reviewed, the final themes are presented as our findings.

5 RESULTS

We present the findings derived from our quantitative and qualitative data analysis. In section 5.1, we show the results of our quantitative data analysis highlighting the overall patterns from the interactions between the LLM-powered agents and participants. As the quantitative data showed high variance, we present the findings of our qualitative analysis in Section 5.2 to Section 5.2.4, to gain further insights into the detailed factors that affected user preference and perceptions towards LLM-powered robots.

5.1 Data from Quantitative Measures

We examined the influence of embodiment on interactions with LLM-powered agents through an analysis of data from our quantitative measures. Figure 3 summarizes significant findings. Overall, embodiment had a significant effect on input prompt length, $F(2, 56) = 14.30, p < .001$. When comparing the input length across embodiment conditions, participants provided significantly longer inputs to the text agent than the voice agent or the robot. Embodiment also had a significant effect on input length within tasks, $F(6, 56) = 4.25, p = .001$. The generation task, in particular, had a significantly longer length of input in the text condition than other embodiment conditions. Finally, embodiment had a significant effect on failures, $F(2, 56) = 55.16, p < .001$. In comparing failures across embodiment conditions, participants encountered the most failures with the voice agent, followed by the robot and text agents. We observed a higher occurrence of failures in the generation task, underscoring the difficulties faced by agents that used voice-based input when confronted with extended input, especially within this

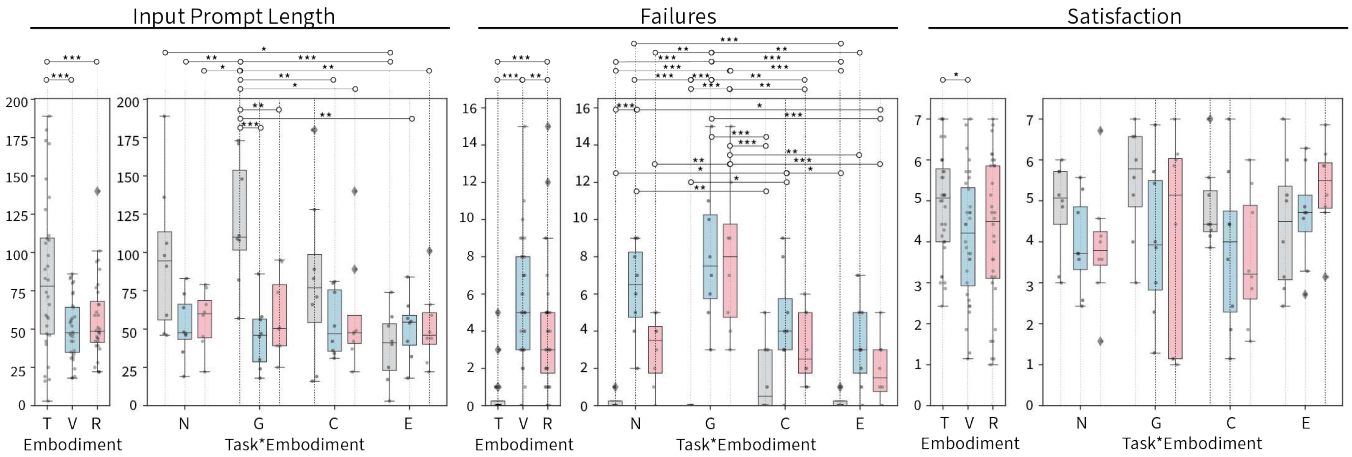

Figure 3: Boxplots with data points overlaid on user satisfaction, length of input prompts, and interaction failures. Embodiment: (T)ext, (V)oice, (R)obot. Tasks: (N)egotiate, (G)enerate, (C)hoose, (E)xecute. Horizontal lines indicate significant pairwise comparisons with Tukey HSD ($p < .05^{*}$, $p < .01^{**}$, $p < .001^{***}$).

specific task context, $F(6, 56) = 5.94, p < .001$. The diverse range of user experiences related to the quality of interactions led to a significant variance in participants' satisfaction when interacting with different agents, $F(2, 56) = 3.81, p = .028$. Participants rated their level of satisfaction with the text agent to be higher than the voice agent and marginally higher than the robotic agent. We attribute these differences in the satisfaction scores to the results of the failures participants experienced with voice-based interaction with the voice and robotic agents. There were no statistically significant differences across embodiments or tasks in other subjective metrics, including anthropomorphism, animacy, likeability, and perceived intelligence, which can be found in the supplementary materials.

5.2 Data from the Qualitative Measures

In this section, we present the findings of our qualitative analysis in the order of tasks in which LLM-powered robots were more preferred by participants, namely: (1) execute; (2) negotiate; (3) choose; and (4) generate. In each task category, we present design themes explaining the positive and negative effects of LLM-powered robot agents, supported by quantitative results.

5.2.1 Execute. Below are themes that emerged in the Execute task.

Conversational Interactions for Effective Learning. Across all the agents, the LLM's capability to facilitate natural conversations while delivering instructions and responses with contextual understanding significantly benefited the participants' engagement in the interaction. For the execution task, the agent received task instructions before engaging with the participant and then responded freely to the participant's requests. All participants frequently sought guidance on how to proceed in the task, thereby leading to concise and clear prompts that were easy for the agent to comprehend and respond to. As shown in Figure 3, the input length of the prompts tended to be shorter in the execution task. Moreover, seven participants expressed satisfaction with the agent's response, the LLM's contextual understanding ability enabled the agent to provide sufficient responses to follow-up questions. *P26: "The robot was able to answer all the spontaneous questions that I had for it, which really surprised me and we were able to have an actual conversation. He's smart enough to teach me!"* Given the seamless communication, there were minimal instances of agents failing to understand and respond logically to requests, as shown in Figure 3.

Robot's Social Aspects Enhancing User Engagement. Six participants expressed a preference for the robot agent over the voice and text agents due to its efficiency in interacting and enriching engagement with the social aspects of the robot. As participants physically prepared drinks while simultaneously seeking instructions from the agent, they expressed that interacting with the robot or voice agent through spoken communication was easier and facilitated multitasking, unlike the text agent, which required them to pause their actions and type queries. *P25: "I could start asking the follow-up question as I was doing a task versus text, I had to finish the whole task and then type the question. I thought it went by a little smoother."* Five participants encountered additional difficulties with the voice agent, struggling to time their prompts with the voice agent, leading to discomfort and reduced interest in engaging with the agent. *P30: "That one [voice agent] for me feels the most choppy and disconnected, so it was hard for me to tell when I could ask something compared to the others [agents]."*

Moreover, four participants noted that the robot's social cues and physical presence enhanced their receptiveness to instructions and task engagement, as it resembled real-life communication. *P26: "When you're able to see Pepper, you can kind of look at the tilted head to understand whether it's like thinking or not. But when you can't see Pepper, it's like, what's going on? Those little things help us communicate."* Four participants also expressed appreciation towards the robot's social cues, such as maintaining eye contact, waiting for task completion, and offering encouragement, as these interactions made participants feel a genuine sense of companionship and support. *P27: "So especially when you're learning, part of the*

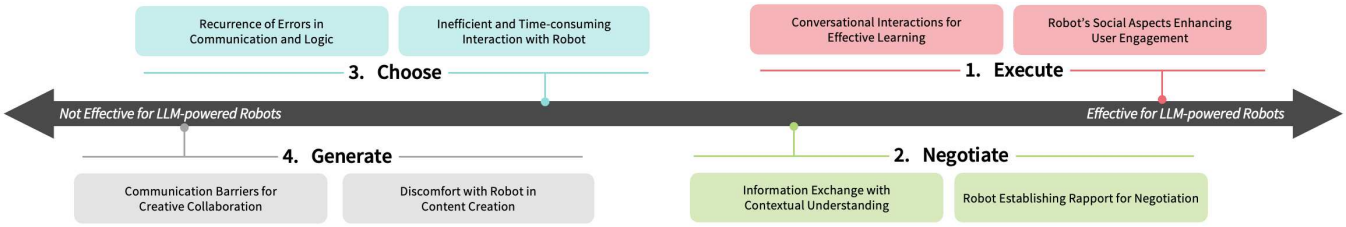

Figure 4: *Summary of Qualitative Findings* — Our findings indicate user preference for LLM-powered robots in the execution and negotiation tasks. These tasks necessitated the establishment of social relationships and rapport, and the robot's social aspects benefited from effective synergy with LLM capabilities. LLM-powered robots were less favored in the choice and generation tasks. In these cases, the robot's interaction medium and its social presence hindered optimal user performance. Additionally, a higher occurrence of technical communication errors contributed to participants' lower preference for robot agents.

learning is from interacting. And that relates to the emotional connections and things that are underneath. So you need the actual robot to do it together, physically engaging." The social cues presented with the robot's social presence increased the participants' focus and immersion in the task, driven by a desire to *P25: "impress the robot, because he is watching me."* Finally for future interactions, five participants envisioned the robot utilizing its arms and body parts for instruction demonstration. Participants explained that they expected the robot to have sophisticated non-verbal cues to match the advanced capabilities of its conversational skills. *P26: "The robot's movements reminded me it was still in development. They were random and didn't have any relation to what it was saying, when the way it talked was such high quality. Made it kind of creepy."*

5.2.2 Negotiate. We found the themes below in the Negotiate task.

Information Exchange with Contextual Understanding. During the negotiation task, participants engaged with the agent to reach a mutual agreement on the price of an item. This negotiation process involved participants posing questions about item specifics, usage history, potential bundle deals, and more. The LLM's ability to understand the context within a conversation, considering the dialogue history to generate coherent responses, was effective in maintaining a seamless, life-like conversation. *P22: "Okay, hold on [robot], first let's sit down and talk about this more. Tell me a little bit more about this bike. Once all my questions are answered sufficiently, then we can start to negotiate."*

Participants' queries for negotiation were generally longer when interacting with the text agent compared to the voice and robot agent, as shown in Figure 3. Five participants explained that it was more convenient to input their questions via text to the agent as opposed to verbally articulating their inquiries. The primary questions posed across the agents were largely similar, with participants posing additional queries based on the agent's response. The text and robot agents had clear distinction when their response was finished, as the text agent displayed a complete response on the screen, while the robot agent indicated completion through non-verbal cues such as putting down its arms, tilting its head, and adjusting its gaze. As illustrated in Figure 3, participants encountered difficulties less frequently when engaging with these agents, whereas the voice agent encountered a higher incidence of failures, as participants had challenges in determining when to speak and whether the agent accurately understood their prompts.

Robot Establishing Rapport for Negotiation. Although the robot agent was not the most convenient to use, four participants expressed that the robot was most effective in establishing the connection and rapport that was required for successful negotiations. Two participants described that building rapport and personal connections with the agent was crucial to resolving the conflicts for negotiation. *P22: "Negotiation starts with building trust, obviously."* The social characteristics, such as the natural language produced by the LLM and the behavioral aspects of gaze, facial expressions, and body movements, contributed to a sense of engaging in a genuine conversation with a social entity. Five participants described that the ability to see and physically interact with the robot agent created a personal interaction atmosphere, enhancing the agent's reliability compared to the voice or text agent that lacked physical social cues. *P17: "I do feel like for me that's important. Just to be able to engage with some visual cues, eyes, you know, a face, that seems to be more appealing, inviting a further engagement. Making a real conversation."* Although the text agent was efficient in providing immediate and informative responses, it was perceived more as a search engine and less as an agent genuinely interested in mediating deals to participants' preferences. This perception made three participants less inclined to negotiate for more expensive items with the text agent. *P5: "Yeah coffee machine no big deal, but the car I wouldn't just negotiate that with the text [agent]. It seems too machinery and sketchy."* Similar to other tasks, the voice agent was least positively perceived among agents for negotiations, as it was challenging to discern intentions and gauge conversational progress during the negotiation.

5.2.3 Choose. We identified the themes below in the Choose task.

Recurrence of Errors in Communication and Logic. During the choosing task, participants chose a final set of items from a list of items based on practicality and preference through discussions with the agents. The discussions required led participants to articulate the reasons for and validate their selections to the agent. Participants also disagreed with the agent's suggestions, prompting them to elaborate on their rationale. Similar to the negotiation task, five participants described that the text agent was the most effective at facilitating accurate and expeditious information exchange

compared to the voice and robot agents. This efficiency resulted in longer prompts in the text agent, as illustrated in Figure 3. Additionally, as participants iterated the item selection criteria and requested validation on the final item selections, the LLM occasionally exhibited errors in logic or inconsistencies with the dialogue history. For instance, one participant described *P24: "He [robot] also had some unusual answers. I asked what I should bring on my ski trip and Pepper said that I could bring a sand baking tray and ski on the tray. I was like, not sure how that would work out. So I felt less engaged in obviously, I felt less of a connection with Pepper in this instance."* In another example, the agent altered its recommendations multiple times or presented conflicting arguments for a single item. Unforeseen failures, such as misinterpreting participants' requests, prematurely responding before participants, and introducing flawed logic in the agent's responses, led to increased failures of the voice and robot agent as illustrated in Figure 3. *P10: "There's often a disconnect between, it [agent] knowing the facts but not knowing that it doesn't make sense."* Such failures were further described by four participants to decrease the satisfaction and motivation to engage with the voice or robot agent.

Inefficient and Time-consuming Interaction with Robot. During the task, participants described that they initially held a general idea of the items they intended to select, and they intended to have specific discussions with the agents to efficiently narrow down their choices. Participants sought details regarding the advantages and disadvantages of these items and validation to determine their inclusion. During this engagement, four participants noted that the conversational interactions necessitated repetitive and overly verbose exchanges with the agent to acquire information equivalent to a "quick search," resulting in an undue amount of time. In contrast, the text-based agent promptly provided the requested information and additionally preserved logs for users to reference when finalizing their decision. *P8: "I appreciated Pepper being all nice, but sometimes it wasn't the exact information I was looking for and there was a lot of fluff. And then I would have to wait for him to finish to ask again. And in the end, I don't even remember what he said! So in those terms, the text was much more efficient."*

5.2.4 Generate. The Generate task included the themes below.

Communication Barriers for Creative Collaboration. The creative nature of the generation task guided participants to devise prompts that were more personal and situation-specific. These prompts included the introduction of character names and attributes, intricate plot developments, and distinctive settings. As a result, participants' prompts tended to be more extensive during the generation task, as shown in Figure 3. A difference in input length appeared between the text agent versus the voice and robot agents, as participants expressed difficulties in verbally expressing their creative thoughts.

These communication difficulties were due to the timely manner of the task. Participants frequently encountered situations where they had a lot to convey but struggled to do so spontaneously in real-time, without excessive pauses or verbosity in response to the robot. Moreover, the collaborative nature of this creative process meant that participants needed additional time to carefully consider how they wanted to incorporate the agent's ideas into their story and shape the subsequent storyline. As a result, seven participants found verbal communication to be less preferable and more challenging compared to using text inputs, where text inputs allowed them to express their ideas in a more organized manner. The communication difficulties led to malfunctions as the agents frequently misinterpreted the user's prompts, overlooked important elements of the participants' instructions, or interrupted the participants. *P7: "But for both cases, the voice and the robot agent. There are a couple of times that they just ignore, or interrupt what you say that makes you more frustrated. It would just start rattling off when I wasn't done talking."* Among all the tasks, the generation task showed the highest number of communication failures, as shown in Figure 3. Thus, six participants perceived the text agent to be more practical in tracking the storyline and avoiding communication errors.

Discomfort with Robot in Content Creation. During the task, four participants expressed discomfort with the robot's social presence when trying to contemplate creative ideas. This discomfort was due to the participants' expectations of the agent being "smart," due to its sophisticated verbal capabilities from the LLM. *P19: "It seems like I am talking to a person, because it [robot] is so smart, and I'm like, oh, they might remember this."* These expectations made the robot's social presence cause pressure on the participants, and even anxiety when they were trying to come up with the next storyline. As the robot would continue to gaze at the user or make subtle movements and facial expressions as it awaited the next prompt, the participants described that this action created a sense of urgency, compelling them to generate their ideas quickly without allowing for thorough reflection. *P6: "I thought more when I was using the text, because it wasn't just off the top of my head. But with the robot, I did just kind of say more random stuff because I felt like I needed to respond right away."* Another participant supported this finding by describing *P19: "I felt the most comfortable with the text event in generating ideas because I had no accountability. Or the robot looking at me in the face. I feel a little bit more embarrassed."*

6 DISCUSSION

In this work, we explored the distinctive design requirements for integrating LLMs with robots. To understand how LLMs should be tailored for robot applications, we conducted a user study involving 32 participants that compared a text, voice, and robot agent across four tasks: execute, negotiate, choose, and generate. Our findings show that the LLM-powered robot elicited user expectations for sophisticated non-verbal cues and was more favored in the Execute and Negotiate tasks, where building connections and engaging in social discussions were crucial. However, LLM-powered robots were less preferred in the Choose and Generate tasks, due to communication difficulties and the potential anxiety during collaboration. Below, we present design implications that address the distinctive design needs for robots utilizing LLMs, as well as the unique design requirements for LLMs intended for use with robots.

6.1 Combining LLM-powered Robots with Non-verbal Interaction Cues

Our findings reveal that interactions with LLM-powered robots established unique expectations for users regarding non-verbal cues. In contrast, users interacting with text and voice-based agents did not actively seek non-verbal cues. These expectations were

not solely shaped by the robot's physical form but were rather a result of the robot's advanced language capabilities powered by the LLM. This sophistication in language abilities led users to anticipate equally sophisticated non-verbal cues from the robot. Therefore, it is recommended that LLM-powered robots explore and incorporate a diverse range of rich non-verbal cues, such as gaze [25], gestures [33, 64], behaviors, and facial expressions [11], during interactions with users. These non-verbal elements should be tailored to match the heightened expectations set by the robot's advanced spoken language capabilities. For instance, combinations of non-verbal cues can be developed to demonstrate appropriate behaviors and explanations in various contexts, such as reactions to different user inputs, respecting user boundaries, and failures in responding to user requests. The alignment of verbal and non-verbal cues can enhance the experience between the user and LLM-powered robots, making the engagement more sophisticated and natural for users.

6.2 Considering Task Characteristics when Utilizing LLMs with Robots

Our study findings indicate that LLM-powered robots exhibited a preference for certain tasks over others due to the unique characteristics of each task. Therefore to effectively utilize LLMs for robots, customization [50, 74] and fine-tuning [37] are crucial for different tasks. While existing state-of-the-art LLMs can be suitable for tasks similar to Execute and Negotiate, for tasks resembling Choose and Generate, LLM adaptation will be required. One method can be fine-tuning LLMs to fit the goal and context of the task, such as simplifying rich social descriptions to enhance efficiency, intuitiveness, and directness. The fine-tuning process can include selecting a pre-trained model, defining task objectives [18], preparing task-specific data [29], configuring fine-tuning parameters [22, 23, 36], training the model, validating and evaluating performance [52], and deploying the fine-tuned model onto the robot.

6.3 Utilizing LLMs for Robot Design

During our study, we observed several design opportunities to leverage LLMs with robots. During the interactions with users, LLMs demonstrated the potential to empower robots to adapt to a broader array of user requests and effectively capture user needs and preferences. These instances emphasize the capacity of LLMs to either substitute or complement traditionally challenging tasks in the realm of robot design and implementation. For instance, during robot application development, significant time is often spent implementing the dialogue system, defining the robot's intent and entity, and training it to handle user requests and communication variability effectively. LLMs can address these challenges by flexibly accommodating task models variations and processing a wide range of inputs. This adaptability can guide robots to offer personalized user experiences through iterative and engaging interactions.

However, it is crucial to acknowledge that integrating LLMs may also introduce risks and errors, such as causing robots to deviate from context or produce hallucination errors. LLM-powered robots in real-world settings may display unexpected behaviors or make statements inconsistent with their intended character, leading to a mismatch between the situational context and the robot's intended personality. Furthermore, as illustrated by instances in our study, LLMs may introduce hallucination errors, leading the robot to provide information that is inaccurate or nonsensical. As a result, LLMs on robots must be viewed as both a feature and a potential challenge, requiring the establishment of appropriate boundaries regarding what LLMs can and cannot achieve. Technical methods such as curated datasets for pre-training [69, 73], program verification [14, 65], human-in-the-loop review [51], fine-tuning, and other measures can be used for LLM action boundaries.

6.4 Limitations and Future Work

Our study has several limitations. First, we chose to compare LLM-powered robots to two other forms in which people interact with LLMs: text agents and voice agents. While this comparison was informative on how people's perceptions of LLM-powered robots differed from other LLM-powered agents, we would have ideally compared the LLM-powered robot to a non-LLM-powered robot. However, it was difficult to specify what a "non-LLM" condition would look like and how such a condition would be implemented. Nonetheless, the lack of comparison against a non-LLM-powered robot limits our ability to study the unique effects of the integration of LLMs in robots. We plan to explore this question in our future work, for example, using a Wizard-of-Oz approach with human operators generating responses or a rule-based approach with scripted dialogues to achieve sophisticated but fixed conversational capabilities for the robot. Second, the quantitative data from subjective measures exhibited high variance, leading to non-significant results in multiple items. This high variability can be attributed to our sample size, which limits the generalizability of our findings. Future work may include larger-scale studies. Third, our minimalist robot design lacked diverse non-verbal behavior, potentially causing users to perceive the three agents as more similar than in real-world scenarios. Future research can explore how LLM-powered robots might use the full range of their embodied capabilities, which could also improve their communication performance with users.

7 CONCLUSION

This research investigates the design requirements for robots connected to LLMs and in tasks where they excel. We compare three LLM-powered agents—text, voice, and robot—across four tasks—generate, negotiate, choose, and execute. Findings reveal that LLM-equipped robots enhance user expectations for non-verbal cues, excel in connection building and deliberation, but face challenges in communication difficulties and creating social pressure. We provide design insights for robots adopting LLMs and LLMs used for robots.

ACKNOWLEDGMENTS

This work was supported by the Sheldon B. and Marianne S. Lubar Professorship, an H.I. Romnes Faculty Fellowship, and the National Science Foundation award (#1925043).

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] Alexander Mois Aroyo, Francesco Rea, and Alessandra Sciutti. 2017. Will You Rely on a Robot to Find a Treasure?. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria)

(HRI '17). Association for Computing Machinery, New York, NY, USA, 71–72. https://doi.org/10.1145/3029798.3038394

[3] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 701–706.

[4] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3 (2011), 41–52.

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. [n. d.]. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. 1, 1 ([n. d.]), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[6] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.

[7] Erik Billing, Julia Rosén, and Maurice Lamb. 2023. Language models for human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction, March 13–16, 2023, Stockholm, Sweden*. ACM Digital Library, 905–906.

[8] Saša Bodiroža. 2017. *Gestures in human-robot interaction*. Ph. D. Dissertation. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät. https://doi.org/10.18452/17705

[9] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. *Springer handbook of robotics* (2016), 1935–1972.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, et al. 2023. FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions. *arXiv preprint arXiv:2308.15214* (2023).

[12] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 293–300.

[13] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[14] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG]

[15] Jessica T DeCuir-Gunby, Patricia L Marshall, and Allison W McCulloch. 2011. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field methods* 23, 2 (2011), 136–155.

[16] Eric Deng, Bilge Mutlu, Maja J Mataric, et al. 2019. Embodiment in socially interactive robots. *Foundations and Trends® in Robotics* 7, 4 (2019), 251–356.

[17] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378 [cs.LG]

[18] Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. OpenAGI: When LLM Meets Domain Experts. arXiv:2304.04370 [cs.AI]

[19] Google. 2023. Google Cloud Services—Speech to text. "Accessed = 09-29-2023".

[20] Guy Hoffman, Oren Zuckerman, Gilad Hirschberger, Michal Luria, and Tal Shani Sherman. 2015. Design and evaluation of a peripheral robotic conversation companion. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 3–10.

[21] Laura Hoffmann and Nicole C. Krämer. 2013. Investigating the effects of physical and virtual embodiment in task-oriented and conversational contexts. *International Journal of Human-Computer Studies* 71, 7 (2013), 763–774. https://doi.org/10.1016/j.ijhcs.2013.04.007

[22] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2790–2799. https://proceedings.mlr.press/v97/houlsby19a.html

[23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *CoRR* abs/2106.09685 (2021). arXiv:2106.09685 https://arxiv.org/abs/2106.09685

[24] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*, Vol. 2. Citeseer.

[25] Bahar Irfan, Sanna-Mari Kuoppamäki, and Gabriel Skantze. 2023. Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adults. https://doi.org/10.21203/rs.3.rs-2884789/v1

[26] Jesin James, Catherine Inez Watson, and Bruce MacDonald. 2018. Artificial empathy in social robots: An analysis of emotions in speech. In *2018 27th IEEE International symposium on robot and human interactive communication (RO-MAN)*. IEEE, 632–637.

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[28] Younbo Jung and Kwan Min Lee. 2004. Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE* 2004 (2004), 80–87.

[29] Weslie Khoo, Long-Jing Hsu, Kyrie Jig Amon, Pranav Vijay Chakilam, Wei-Chu Chen, Zachary Kaufman, Agness Lungu, Hiroki Sato, Erin Seliger, Manasi Swaminathan, Katherine M. Tsui, David J. Crandall, and Selma Sabanović. 2023. Spill the Tea: When Robot Conversation Agents Support Well-Being for Older Adults. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 178–182. https://doi.org/10.1145/3568294.3580067

[30] Krishna Kodur, Manizheh Zand, Matthew Tognotti, Cinthya Jauregui, and Maria Kyrarini. 2023. Structured and Unstructured Speech2Action Frameworks for Human-Robot Collaboration: A User Study. (2023).

[31] Guy Laban, Jean-Noël George, Val Morrison, and Emily S Cross. 2020. Tell me more! Assessing interactions with social robots from speech. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2020), 136–159.

[32] Christine P Lee, Bengisu Cagiltay, and Bilge Mutlu. 2022. The unboxing experience: Exploration and design of initial interactions between children and social robots. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–14.

[33] Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models. arXiv:2308.16529 [cs.RO]

[34] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5 (2013), 291–308.

[35] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[36] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems* 35 (2022), 1950–1965.

[37] Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, and Xiang Li. 2024. Tailoring Large Language Models to Radiology: A Preliminary Approach to LLM Adaptation for a Highly Specialized Domain. In *Machine Learning in Medical Imaging*, Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang (Eds.). Springer Nature Switzerland, Cham, 464–473.

[38] Arnold M Lund. 2001. Measuring usability with the use questionnaire12. *Usability interface* 8, 2 (2001), 3–6.

[39] Michal Luria, Jodi Forlizzi, and Jessica Hodgins. 2018. The effects of eye design on the perception of social robots. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1032–1037.

[40] Michal Luria, Guy Hoffman, Benny Megidish, Oren Zuckerman, and Sung Park. 2016. Designing Vyo, a robotic Smart Home assistant: Bridging the gap between device and social agent. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1019–1025.

[41] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing social robot, screen and voice interfaces for smart-home control. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 580–628.

[42] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 633–644.

[43] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3 (11 2019), 1–23. https://doi.org/10.1145/3359174

[44] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

[45] Jonathan Mumm and Bilge Mutlu. 2011. Human-Robot Proxemics: Physical and Psychological Distancing in Human-Robot Interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 331–338. https://doi.org/10.1145/1957656.1957786

[46] Bilge Mutlu. 2021. The virtual and the physical: two frames of mind. *iScience* 24, 2 (2021), 101965. https://doi.org/10.1016/j.isci.2020.101965

[47] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 1–33.

[48] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica Hodgins, and Sara Kiesler. 2006. Task structure and user attributes as elements of human-robot interaction design. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 74–79.

[49] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 61–68.

[50] Teresa Onorati, Álvaro Castro-González, Javier Cruz del Valle, Paloma Díaz, and José Carlos Castillo. 2023. Creating Personalized Verbal Human-Robot Interactions Using LLM with the Robot Mini. In *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)*, José Bravo and Gabriel Urzáiz (Eds.). Springer Nature Switzerland, Cham, 148–159.

[51] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

[52] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv:2302.12813 [cs.CL]

[53] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 145–152.

[54] Samantha Reig, Elizabeth J Carter, Terrence Fong, Aaron Steinfeld, and Jodi Forlizzi. 2022. Perceptions of explicitly vs. implicitly relayed commands between a robot and smart speaker. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1012–1016.

[55] Samantha Reig, Jodi Forlizzi, and Aaron Steinfeld. 2019. Leveraging robot embodiment to facilitate trust and smoothness. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 742–744.

[56] Aldebaran Robotics. 2023. Animated Speech. "Accessed = 09-29-2023".

[57] Aldebaran Robotics. 2023. Audio Device API. "Accessed = 09-29-2023".

[58] Soft Bank Robotics. 2023. Pepper Robot. "Accessed = 09-29-2023".

[59] Eduardo Rodriguez-Lizundia, Samuel Marcos, Eduardo Zalama, Jaime Gómez-García-Bermejo, and Alfonso Gordaliza. 2015. A bellboy robot: Study of the effects of robot behaviour on user engagement and comfort. *International Journal of Human-Computer Studies* 82 (2015), 83–95. https://doi.org/10.1016/j.ijhcs.2015.06.001

[60] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4 (2012), 201–217.

[61] Guido Schillaci, Saša Bodiroža, and Verena Vanessa Hafner. 2013. Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics* 5 (2013), 139–152.

[62] Elena Márquez Segura, Michael Kriegel, Ruth Aylett, Amol Deshmukh, and Henriette Cramer. 2012. How do you like me in this: User embodiment preferences for companion agents. In *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12*. Springer, 112–125.

[63] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. 2015. Poor Thing! Would You Feel Sorry for a Simulated Robot? A Comparison of Empathy toward a Physical and a Simulated Robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 125–132. https://doi.org/10.1145/2696454.2696471

[64] Gabriel J Serfaty, Virgil O Barnard, and Joseph P Salisbury. 2023. 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* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 13, 3 pages. https://doi.org/10.1145/3586182.3616623

[65] Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate & Rank: A Multi-task Framework for Math Word Problems. arXiv:2109.03034 [cs.CL]

[66] 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.

[67] Leila Takayama, Victoria Groom, and Clifford Nass. 2009. I'm Sorry, Dave: I'm Afraid i Won't Do That: Social Aspects of Human-Agent Conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2099–2108. https://doi.org/10.1145/1518701.1519021

[68] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503* (2021).

[69] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

[70] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res* 2 (2023), 20.

[71] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 117–122.

[72] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2007. Embodiment and human-robot interaction: A task-based perspective. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 872–877.

[73] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560 [cs.CL]

[74] Takato Yamazaki, Katsumasa Yoshikawa, Toshiki Kawamoto, Tomoya Mizumoto, Masaya Ohagi, and Toshinori Sato. 2023. Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models. *Advanced Robotics* 37, 21 (2023), 1364–1381.

[75] Yang Ye, Hengxu You, and Jing Du. 2023. Improved Trust in Human-Robot Collaboration With ChatGPT. *IEEE Access* 11 (2023), 55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111

[76] Tom Ziemke. 2013. What's that thing called embodiment? In *Proceedings of the 25th Annual Cognitive Science Society*. Psychology Press, 1305–1310.

[77] Zoom. 2023. Video Conferencing Platform. "Accessed = 09-29-2023".