Structured and Unstructured Speech2Action Frameworks for Human-Robot Collaboration: A User Study

Krishna Kodur   Manizheh Zand   Matthew Tognotti   Cinthya Jauregui   Maria Kyrarini

***Abstract*—This research delves into user preferences concerning structured (the subject follows an exact script to command the robot) and unstructured (the subject commands the robot in a conversational way) robot interaction through natural spoken language. Data was gathered from 30 adult participants who completed two distinct tasks involving both structured and unstructured commands. The study examines correlations between robot errors and user perceptions, as well as how past or present failures impact participants' perception of robot utility. Three hypotheses are formulated, and the paper offers a comprehensive overview of the study's aims, methodologies, and principal findings, which were ascertained using paired t-Test and Kendall-Tau correlations. The study indicates that participants showed a preference for the unstructured task in contrast to the structured one. Analysis of the data revealed interesting correlations between the user perception of the robot and the robot errors.**

***Index Terms*—Structured Speech, Unstructured Speech, Human-Robot Interaction, Voice commands, Human-Robot Interface, User Preferences, Natural Language Processing, Large Language Models**

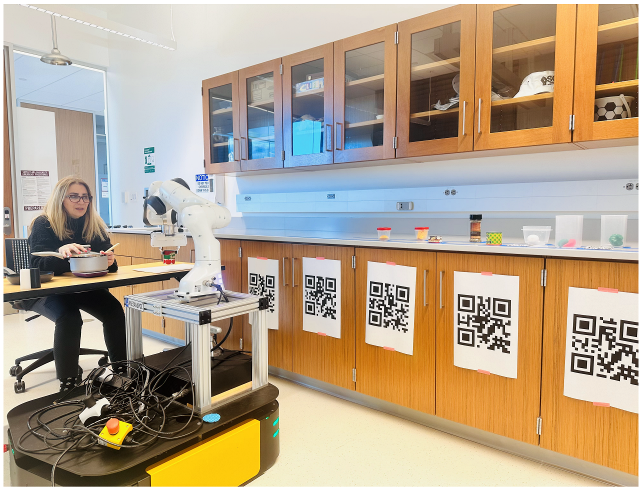

Fig. 1: Setup for collaborative cooking scenario with a robot; Robot is delivering tomato sauce

I. Introduction

In recent years, robots are on the rise in our homes. According to the International Federation of Robotics Report, 2019 [1], there is an uptick in the adoption of robots for households. These robots can be defined as household robots that can be deployed at homes to perform routine tasks, e.g., vacuuming the floor, fetching objects, assisting with cooking, etc. This crucial integration of robots in households necessitates the development of efficient modes of communication to interact with them. Robot manufacturers currently offer users various graphical interfaces, such as website dashboards or mobile apps, to operate their robots. However, these interfaces might not align well with natural modes of communication (e.g., speech) and could present accessibility challenges for differently-abled users, making them potentially unsuitable for this user group [2], [3]. Speech-based communication emerges as a viable and natural approach to make the robots more accessible and easier to use at home, enabling them to continually learn from the inputs they receive.

One approach to human-robot interaction via speech is a predefined vocabulary that the user can employ to communicate with the robot, and the robot is able to understand and perform the requested actions [4]–[7]. This type of interaction is *structured*, as the commands and the robot actions are well-mapped to each other. Taking advantage of the recent breakthrough of Natural Language Processing (NLP) with ChatGPT, humans, and robots can now communicate in an *unstructured* fashion [8]. This approach may feel seamless for humans, but it can be more challenging for the robots to recognize the required actions [9], [10]. Our research assesses user experiences with robots by analyzing the user study conducted in a laboratory, which simulates a home setting. The user study compares structured and unstructured communication modes between humans and mobile manipulators. Nowadays, robotic cooking assistants are gaining popularity [11]–[13]. Therefore, as an interactive scenario, a collaborative cooking task is selected, as illustrated in Figure 1.

The contribution of this paper lies in understanding user preferences for speech-based interaction with a real robot in a collaborative cooking scenario, which includes robot failures. Therefore, we defined three hypotheses that are designed to facilitate an understanding of these preferences, as elaborated in the following subsection I-A. To the best knowledge of

Krishna Kodur, Manizheh Zand, Matthew Tognotti, and Dr. Maria Kyrarini are with the Department of Electrical & Computer Engineering at Santa Clara University, Santa Clara, CA 95053 USA (email: {kkodur, mzand, mtognotti, mkyrarini}@scu.edu)

Cinthya Jauregui is with the Department of Engineering Management & Leadership at Santa Clara University, Santa Clara, CA 95053 USA (email: cjauregui@scu.edu)

the authors, this is the first study that compares the two modes of communication while including an interactive and collaborative robotic scenario. Additionally, in accordance with the guidelines set forth by the Institutional Review Board (IRB) protocol ID 23-02-1902, the dataset from the study is shared through our lab's official website (Dataset: https://sites.google.com/view/hmi2lab/datasets). The outcomes of this study will offer valuable insights to inform the design of forthcoming human-robot interaction systems.

A. *Hypothesis*

The primary objective of this research is to determine if individuals prefer structured or unstructured spoken language to instruct the robot. Additionally, the study aims to investigate how robot failures impact the individual's perception, specifically in relation to structured and unstructured methods. To this end, the paper presents three hypotheses:

1) **Hypothesis 1:** Individuals exposed to unstructured robot interaction via spoken language will demonstrate a higher preference for this mode of interaction compared to structured robot interaction.
2) **Hypothesis 2:** The individual's perception of robots will be negatively affected when they encounter robot failures during the interaction, as opposed to instances without or with minimal failures.
3) **Hypothesis 3:** The individual's preferred method of instructing the robot, whether structured or unstructured, will be influenced by their previous experiences with robot failures during the interaction, based on the respective method.

The following sections of this paper are structured as follows: Section II outlines the existing literature on different methods of human-robot communication involving speech or text interfaces. In Section III, an overview of the robotic system used to examine interactions between humans and robots is provided, including details about the experimental scenarios and protocols. Section IV conducts a comprehensive analysis of the proposed hypotheses, determining their acceptance or rejection based on the gathered data. The implications of the hypothesis analysis results are discussed in depth in Section V. Finally, Section VI concludes the paper by summarizing the findings and addressing potential avenues for future research.

II. Related Work

Human-Robot Interaction (HRI) involves various communication modes to facilitate seamless interaction between humans and robots. Among the primary communication modes in HRI are Graphical User Interfaces (GUIs) with buttons and text, gestures, and speech interfaces. These interfaces have been extensively explored in the field of robotics [14]–[21]. Speech interaction with robots can be of two types: structured or unstructured. Structured speech entails clear, organized, and specific instructions, with clear intent but with limited use of vocabulary. While unstructured speech refers to more spontaneous and informal language, which encompasses a larger vocabulary; however, the intent can sometimes be hard to decipher. To understand the significance of speech, it is crucial to compare various other communication modes. Strazdas et al. [16] conducted a Wizard-of-Oz study with 36 participants to analyze how the users would interact with a robot using unstructured speech and gestures. The authors teleoperated the robot behind the scenes whenever a user gives either a speech or gesture command to perform a task, e.g., pick and place an object. The subjects were given the option of interacting via speech, gesture, or a combination of both; 97.2% of the participants used speech at least once to command the robot. This percentage highlights the importance of speech in human-robot interaction.

Recent advancements in Large Language Models (LLMs) have paved the way for robots capable of understanding and processing both structured and unstructured text [8], [22]. In the current literature, there are many robotics systems proposed that use structured speech as their communication mode. Chen et al. [17] proposed a human-robot collaboration in an industrial setting where the users can command the robot using structured speech and gestures. However, the vocabulary used is very limited such as “start”, “stop”, “go home”, etc. Another speech-based robot interaction framework was proposed by Giorgi et al. [18]. The authors presented a novel approach for acquiring high-level task-learning capabilities in robots, illustrated by the task of “making tea”. However, structured language is utilized to facilitate communication between the robot and humans. When given a speech command by the user, such as “make me a tea”, it was represented as an array of low-level actions, thus enabling the robot to execute complex tasks sequentially by combining individual actions, e.g., ”mug_grasp_lift_table_drop,” ”bottle_grab_lift_mug_pour,” and ”teabag_grab_pickup_mug_throw” for the task of making tea. Each low-level action, such as grab, etc, is preprogrammed. Instead of representing a task as an array of low-level actions, Shao et al. [23] introduced a novel framework designed to enable a robot‘s execution of diverse object manipulation tasks, illustrated by an example “Put a cup in front of the bowl”. The proposed framework used Bidirectional Encoder Representations from Transformers (BERT) [24], an LLM in conjunction with a CNN-based deep neural network called ResNet [25] to generate robot trajectories. The framework is fed with both an instruction text and an image depicting the initial scene. In response, the model generates a robot motion trajectory to successfully accomplish the given task. However, it should be noted that Shao et al.'s framework was constrained by the use of pre-defined language templates sourced from the Something-Something dataset [26].

A more advanced mode of interacting with robots is by using unstructured text inputs, either using a mobile app or a website. While it is different from users employing unstructured speech for interaction, they share a common characteristic in terms of communicating intent through unstructured means. One such example is the SayCan robot, developed by Ahn et al. [19], which used an LLM to receive

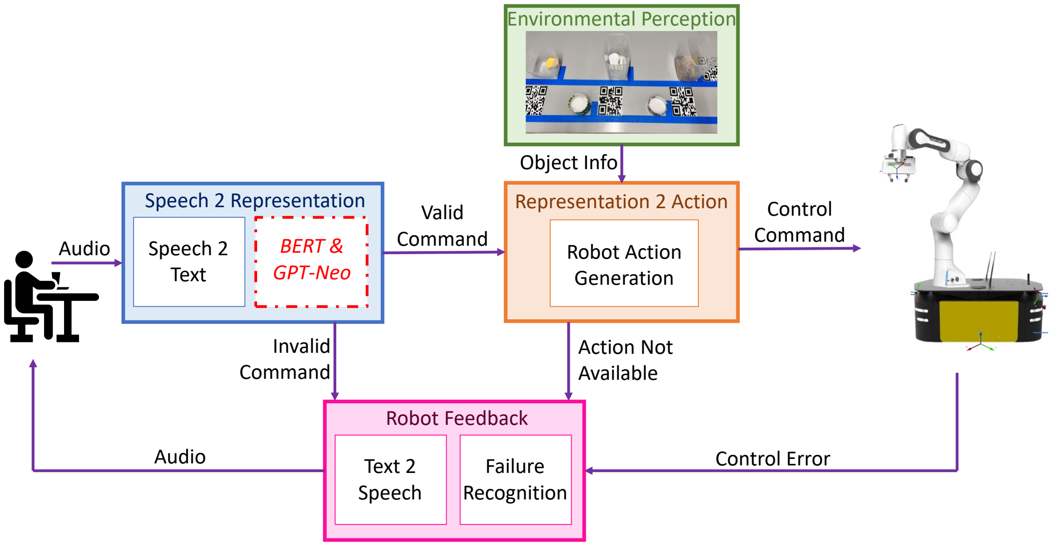

Fig. 2: Overview of the proposed system for unstructured and structured human-robot communication. Both modes of communication consist of all the blocks, with the exception of the block "BERT & GPT-Neo", which is only part of the unstructured communication.

high-level tasks from users. SayCan developed a kitchen robot that can fetch objects such as Coca-Cola. This robot received high-level tasks in a text form from the user through a GUI, such as "I spilled a coke; can you bring me something to clean it up?". The Pathways Language Model (PaLM) [22], an LLM developed by Google, was used to train and generate a series of instructions that would break down the high-level task into low-level tasks so that the robot can perform, such as 1) Find Sponge, 2) Pick up Sponge, 3) Bring it to the user, 4) Done. The users can use both structured and unstructured text data to interact with the robot. However, despite the robot‘s capabilities, user preference data on the type of text communication, such as structured or unstructured, is currently lacking. In another study conducted by Ye et al. [20], the authors used a popular LLM called ChatGPT and fine-tuned it. RoboGPT was deployed in an industrial assembly process inside a Virtual Reality (VR) simulation. The users were able to command the robot using their unstructured speech. Their speech was converted to text and then fed into RoboGPT. Based on RoboGPT‘s intent inference, RoboGPT controlled the robot to perform the task. Data was collected from 15 participants on how they perceived the system. The authors concluded that the integration of ChatGPT in robots has shown a notable increase in trust during human-robot collaborative interactions. These findings are of interest; however, additional research is needed to validate them in real-world human-robot collaborative scenarios.

Therefore, the question that emerges pertains to the type of speech communication preferred by users, structured or unstructured, while interacting with the robot in realistic scenarios. There is a lack of research exploring the use of speech as a communication mode and investigating the differences between structured and unstructured speech in the context of human-robot interaction. This paper studies the preferred speech method of interaction between humans and robots and specifically how humans perceive robots based on trust and ease of use in relation to the structured or unstructured type of speech interaction while considering the effect the robot errors.

III. Experimental Methodology

A. *Overview of the Robotic System*

The primary objective of this study is to examine an individual‘s inclination towards either structured or unstructured speech in human-robot interactions within a natural setting. With this in mind, a kitchen setting is created in the lab. The individual sits behind a table and instructs the robot using speech to fetch items for them, and then the robot has to perform that task, just as an ideal cooking helper (sous chef) robot would be expected to perform in real life.

Hardware Setup: The robot used for the interaction is named the "SousChef Robot" and, in short, "SousChef". It is named as such because it can interact with the users as a sous chef via normal speech using Artificial Intelligence models. SousChef is a mobile manipulator that can fetch ingredients related to cooking, such as pasta, tomato sauce, green beans, butter, etc. SousChef comprises two robotic platforms integrated together: the 7-Degrees of Freedom Franka Emika Panda robotic manipulator (Panda) and the Clearpath Ridgeback mobile base robot (Ridgeback). Panda is equipped with seven joints with torque sensors at each joint and a payload capacity of 3 kg. Panda is mounted on the Ridgeback, which is an omnidirectional Mecanum wheel system capable of carrying payloads of up to 100 kg. SousChef is equipped with two cameras, Intel Realsense D455 (D455) and Intel Realsense D405 (D405). D455, placed on top of the Ridgeback, is used for the navigation of the Ridgeback, while D405, placed on top of the end-effector of Panda, is used to recognize objects in the environment. The Robot Operating System (ROS) framework is employed to orchestrate robot motion, facilitating synchronized operation among multiple robots and cameras.

System Overview: Figure 2 shows an overview of the SousChef's system architecture. The users can interact with the SousChef to fetch objects in two modes, structured or unstructured speech. The SousChef system depicts the holistic view of human-robot interaction using speech in both modes and consists of the following four modules: Speech 2 Representation, Environmental Perception, Representation 2 Action, and Robot Feedback.

Speech to Representation (S2R) Module: The S2R module initially employs Google Cloud Speech-to-Text, which converts the spoken words into sentences. In the structured mode, the user is required to speak with the following structured sentence: "Give me the [object name]", where [object name] is an object in the environment. The module identifies a command as valid if it adheres to the defined structure or as invalid if the command deviates from the defined structure. In unstructured interactions, in which users have the freedom to instruct the robot in their own words, natural language understanding models are employed. To achieve this, Bidirectional Encoder Representation from Transformers (BERT) [27] is first used to categorize textual content into two distinct classes; valid commands related to cooking, encompassing activities such as object fetching, and invalid commands aligned with a general discourse on topics other than cooking. The valid commands are subsequently processed by Generative Pretrained Transformer (GPT) Neo [28], which transforms the sentence into a properly structured command. BERT and GPT Neo are retrained on our "Collaborative Cooking Dataset" [9], which contains speech and corresponding text data of how users would interact with the robot, asking the robot to fetch objects, place objects from one location to the other, setting timers, etc. The valid commands from both the unstructured and structured interactions are sent to the Representation To Action module, and the invalid commands are forwarded to the Robot Feedback module.

Environmental Perception Module: For mobile manipulation of the objects in the scene, QR codes [29] are deployed that include information about object ID and location. Ridgeback uses QR codes on the cabinets, as shown in Figure 1, for navigation and localization, while the QR codes next to the objects are used by Panda for object recognition. It is necessary to note that the focus of this research is not on advancing robotic vision; therefore, QR codes are selected as a well-established and accurate approach to object recognition [30].

Representation to Action (R2A) Module: The R2A Module receives the valid commands and the information about the objects within the scene as inputs and subsequently generates the robot control commands for both the Ridgeback and the Panda. SousChef is a fetch robot, and it can fetch the items present in the environment. If the user requests an object that is present in the scene, R2A generates the robotic commands that enable the robot to move to the required location, pick the required object, move it near the user, and place it on the user's table (see Figure 1). Additionally, R2A can recognize similarly named objects present in the environment. For example, users usually ask for bell pepper as pepper. At that time, the module finds the nearest possible item in the environment by using the Jaro-Winkler similarity [31]. If the user requests an object that is not present in the scene, then the action is considered unavailable, and the robot feedback module is notified.

Robot Feedback Module: The robot feedback module is responsible for communicating with the user regarding potential issues that arise during the interaction. For example, if the user requests an object that is not present (therefore, the action is not available), the robot responds by saying, "The item is not present" by using Google Cloud Text-To-Speech. If the user provides an invalid command (based on the output of the S2R module) by asking for a task that the robot is not programmed to do, the robot responds by saying, "It's a tough task. Can you ask me for something that I can do?". Although the system is designed to minimize the errors to the maximum possible extent, errors do occur during human-robot interaction or during the execution phase, where the robot fetches items for the user. The robot feedback module recognizes the following failures; (1) Grip & Slide error - It happens when Panda fails to grip the object properly, and instead, it keeps pushing the object from the side, so the object keeps sliding, (2) Grip & Miss error - This error happens when Panda misses gripping the object and does not touch the object at all, (3) Grip & Drop Error - This error results when Panda grabs the object but drops it after gripping, and (4) Grip & Drag Error - This error occurs when Panda grabs the object but does not lift the object to a proper height, and as a result, the object drags on the counter. If one of these errors is detected while performing an action commanded by the user, SousChef returns to its home pose. The researchers then reposition the object to its original position, and the action is restarted.

B. Experiment Scenario

The purpose of the research is to study how individuals interact with robots at home to run their Activities of Daily Living (ADL). As we are interested in developing an accessible system for both body-abled and people with limited lower-body mobility, the user sits on a chair beyond a table with minimal movement, as shown in Figure 1. This research is primarily focused on preparing meals at home, such as cooking pasta. The table represents the cooking area where the user has access to the stove and water, and some utensils, such as pots and pans. Items provided on the counter, out of the user's reach, are bell peppers, butter, carrots, cheese, chili, corn, garlic, green beans, mushrooms, pasta, and tomato sauce. Since the focus of the study is to identify the preferred method of individuals interacting with the robot, the study is split into two separate tasks. In one task, the individual commands the robot via structured speech, whereas in the other task, the individual commands the robot using unstructured speech, naturally, without memorizing the sentences of how to interact with the robot.

C. Experiment Protocol

In order to investigate the user's speech interaction preferences with a robot, the experimental protocol is designed to systematically evaluate the user's perception of the SousChef robot system. System Usability Scale (SUS) [32] and Human-Robot Collaboration Questionnaire (HRCQ) surveys are used to evaluate the user's perception. The SUS survey is a widely-used survey designed to assess the perceived usability and user-friendliness of a product or system. The HRCQ, shown in Table I, is a custom questionnaire inspired by [33], [34], designed to get insights into various HRI aspects of the system.

Recruiting: Thirty adult participants, on a voluntary basis, were recruited from faculty, students, visitors, and staff from the School of Engineering at Santa Clara University. Of the 30 participants, 8 (26.6% of participants) were female, and 22 were male. This ratio between females and males is similar to the ratio at the School of Engineering at Santa Clara University. The advertisement of the study was conducted via e-mail. Participants who showed interest in the study booked an appointment. A letter of the consent form was emailed to the participant before their acceptance to participate in the study, which indicated the duration of the study, one hour, and how the participant's information would be kept confidential. The study has been approved by Santa Clara University's IRB with protocol number 23-02-1902.

Order of tasks on the day of the study: The SousChef robot system is located at Santa Clara University in the Human-Robot Interaction and Innovation (HMI^2) lab. Approximately 20 minutes prior to the participant's arrival, the system is initiated to bring it up and running. The system, as shown in Figure 1, consists of a table that the subject sits behind, away from the robot and the ingredients. The SousChef's Ridgeback is at the starting default position (home position). Following IRB protocol number 23-02-1902, to create a safe environment for the user in the lab, the speed of the robot is reduced to a maximum of 0.05 meters per second. Upon the subject's arrival, they are asked to sign a consent form and are optionally provided with a demographic survey. A researcher then explains the upcoming data collection process, and the participant is instructed to sit on a chair during the entire study.

To minimize bias between the two modes of communication, structured and unstructured, we followed a counter-balanced approach, i.e., the order of the tasks (structured vs unstructured) is switched for subsequent subjects. For instance, if a participant started first with the structured task and then did the unstructured task, then the next participant will start with the unstructured task and complete the study with the structured task.

TABLE I: The Likert Scale Statements for the Human-Robot Collaboration Questionnaire.

	Likert Scale Based Statements
Perceived Usefulness	I accomplished the given tasks rapidly.
	I accomplished the given tasks successfully.
Perceived Safety and Trust	The robot's actions were predictable.
	I felt safe using the robot.
	I trusted the robot's suggestions.
Perceived Ease of Use	I found the robot easy to use.
	The robot learned how to assist me.
	The robot met my expectations.
Perceived Interaction	I had to learn more about robots in order to be able to interact with the system.
	I felt my voice volume was normal.
	I had to speak slowly to interact with the robot.
Ethical Considerations	It is acceptable for the robot to have much information about the user.
	I am concerned about my privacy when using the robot.
	I should have full control of when and how the robot will assist me.

After the briefing, the study started. A microphone is placed on the table next to the subject that can be turned on while giving the command and turned off while not in use. Upon receiving a command via the microphone, the SousChef will fetch the item and place it on the table next to the user. This process is repeated until the participant feels that the cooking has been completed. The participant then completes the SUS and HRCQ survey. After the first task is completed, the process will be repeated for the second task. After the second task is completed and the participant completes another SUS and HRCQ survey, the visit is concluded.

While conducting the study, one of the researchers constantly follows the robot to ensure the subject's safety by holding an emergency button. The other researcher monitors the data collection and intervenes in case of robot failures.

Reporting the results: Data analysis is based on SUS and HRCQ surveys. HRCQ covers the human-robot interaction-based metrics such as Perceived Usefulness, Perceived Safety and Trust, Perceived Ease of Use, Perceived Interaction, and Ethical Considerations on a Likert scale from 1-5 (where "1" represents strongly disagree and "5" represents strongly agree), as shown in Table I. After the Likert Scale statements, the following optional open questions are asked: (1) "What additional functionalities should the robot have?", (2) "What did you like about the robotic system?" (3) "What frustrated you about the robotic system?" and (4) " Please provide any additional comments/feedback for the robotic system".

All the collected data has undergone anonymization procedures to ensure the preservation of the privacy of the participants. The dataset is accessible through our project's official website, as described in Section I. The data includes a demo video, SUS questionnaires, HRCQ surveys, and demographic information and code for the SousChef robot strategically provided to stimulate interest in Human-Robot Interaction (HRI) studies. The dataset is then used for our hypotheses analysis, as shown in the next Section.

IV. Statistical Analysis and Results

Within this Section, a comprehensive statistical analysis is conducted to examine the hypotheses outlined in Section I-A.

Hypothesis 1: To ascertain the preference for unstructured robot interaction through spoken language over structured robot interaction among participants, the t-Test [35] is employed. Prior to conducting hypothesis analysis using a t-test, it is imperative to ensure that the metric under consideration, which compares structured and unstructured modes, adheres to a normal distribution. The chosen metric for evaluation is the System Usability Scale (SUS) score, as it provides a comprehensive insight into user perceptions.

The Shapiro-Wilk test [36] is employed as a means to ascertain whether the collected SUS scores are drawn from a normal distribution. If the p-value resulting from the Shapiro-Wilk test surpasses the significance threshold of 0.05, it establishes statistically substantial evidence indicating that the distribution of the data aligns with a normal distribution. For the SUS scores for the structured task, the computed p-value is 0.9832, while for the unstructured task, the p-value is 0.1481. The p-value for both structured and unstructured SUS scores exceeds 0.05, indicating that both sets are likely drawn from a normal distribution. This conclusion is further substantiated by the visual depiction of the SUS scores through violin plots, where the characteristic bell curve shape associated with a normal distribution is evident, as illustrated in Figure 3. It can be seen by Figure 3 that the 50^{th} percentile of the unstructured tasks is much higher than the 50^{th} percentile of the structured tasks.

The paired t-test is a statistical method and is used to compare the means of two related groups, such as before-and-after measurements or matched pairs. It assesses whether there is a significant difference between the means while accounting for the inherent correlation or pairing between the observations. By calculating the t-statistic based on the differences between paired observations and their standard deviation, the paired t-test helps determine whether the observed differences are likely due to chance or if they reflect a true underlying change.

In pursuit of the hypothesis's objectives, the subsequent Null and Alternative hypotheses are formulated as follows:

- Null Hypothesis (H_0): There exists an absence of statistically significant evidence indicating the user's preference for unstructured modes of communication over structured alternatives. There is no difference between underlying distributions of structured and unstructured SUS scores.
- Alternative Hypothesis (H_a): Compelling statistical evidence indicates that users exhibit a preference for unstructured modes of communication over structured ones. The unstructured SUS scores tend to be greater than the structured SUS scores.

In substantiating the null hypothesis, the SUS scores for both unstructured and structured communication modes are computed, followed by the application of the paired t-test. It is noteworthy that the mean SUS score for unstructured communication mode was determined to be 79.83 ± 13.86, while the mean SUS score for structured communication mode was calculated at 73.16 ± 13.64. The resulting p-value, calculated as 0.0016, was found to be less than the predetermined significance level of 0.05, thus leading to the confirmation of the alternate hypothesis. Furthermore, observed SUS averages serve to provide additional support for the hypothesis in question.

Hypothesis 2: In order to examine whether individual's perceptions of robots are adversely influenced by encounters with robot failures, particularly instances where the robot struggled to comprehend the user's intentions or faced challenges in the successful retrieval and delivery of items during interactions, the Kendall's Tau Correlation method was employed [37]. This methodology assesses the extent of ranking similarity and is utilized to determine statistically significant correlations between the occurrence of robot error rates and SUS scores, as well as specific questions from the HRCQ surveys detailed in Section III-C. The robot error rate can be defined as the total number of robot errors divided by the total number of user commands.

To test whether there is a correlation between encounters with robot failures and an individual's perception of the robot using the Kendall-Tau method, the subsequent Null and Alternative hypotheses are formulated as follows:

- Null Hypothesis (H_0): There is no statistically significant correlation between encounters with robot error rate and individuals' perceptions of robots, as measured by SUS scores or specific HRCQ questions.
- Alternative Hypothesis (H_a): There is a statistically significant correlation between encounters with robot error rate and individuals' perceptions of robots, as measured by SUS scores or specific HRCQ questions.

To understand how individuals perceive robot errors, Table II summarizes the correlation between SUS and robot error rate considering all the possible scenarios: All structured and unstructured tasks (All Tasks (S+U)), structured only tasks (S Task), unstructured Only Tasks (U Task), all first tasks which could be either structure or unstructured (Task-1 (S or U)), second tasks which could be either structure or unstructured (Task-2 (S or U)), first tasks that are only structured (Task-1 = S Task), first tasks that are only unstructured (Task-1 = U Task), second tasks that are only structured ((Task-2 = S Task)), and second tasks that are only unstructured (Task-2 = U Task). According to Schober et al., [38], correlation is considered statistically significant and moderate if and only if the correlation is greater than or equal to 0.26 and the p-value is less than 0.05. As shown in Table II, there are no correlations between SUS and robot error rate. To expand the relationship between user feedback and robot errors, metrics from the HRCQ have been compared extensively to the robot error rate.

As shown in Table II, it is found that there is a negative correlation between "I felt safe using the robot" during the

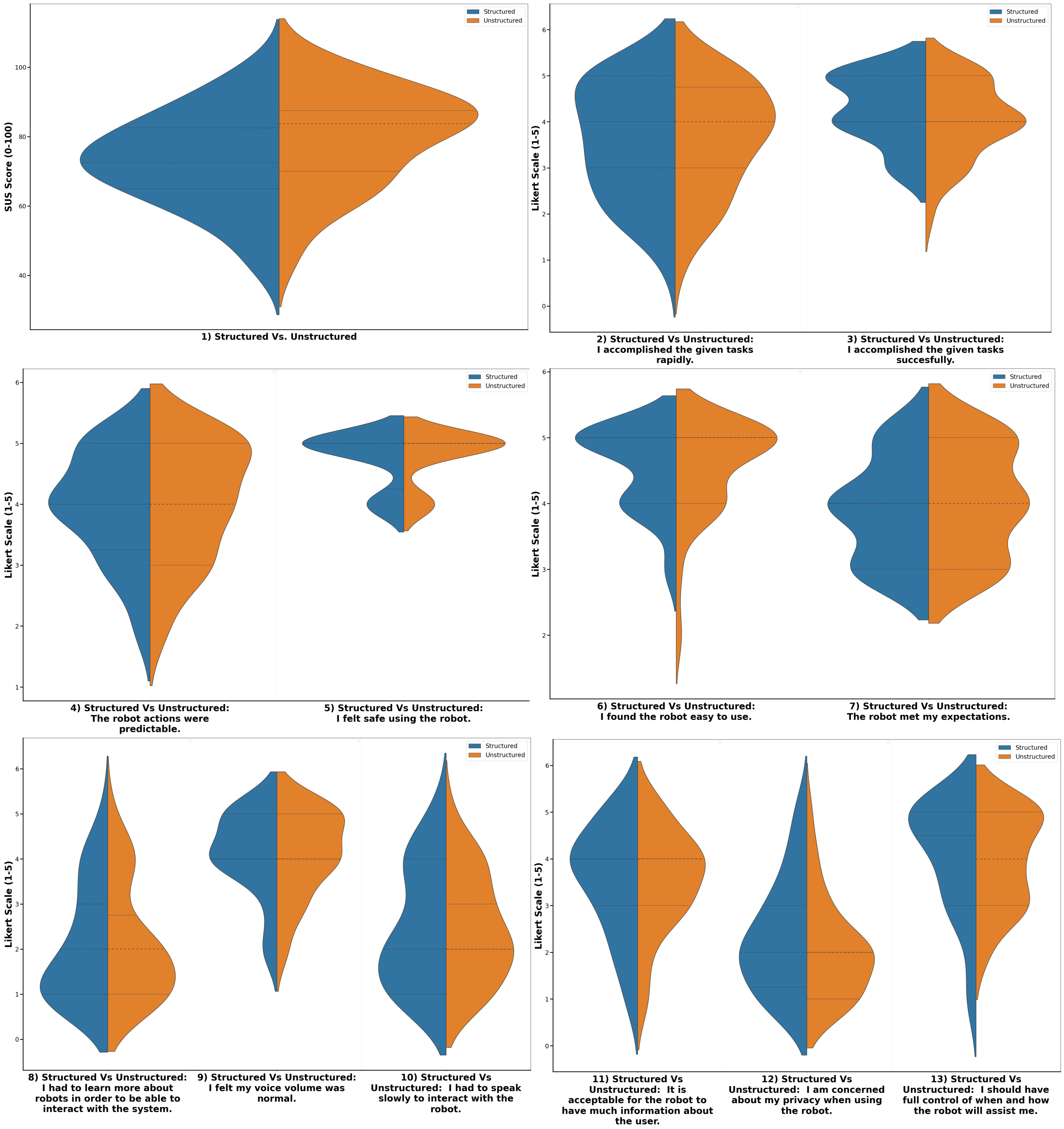

Fig. 3: Violin plots for SUS and HRCQ survey results for both structured and unstructured tasks. Data from the structured task is visualized in blue on the left half, while the data from the unstructured task is depicted in orange on the right. Dotted lines indicating the 25^{th}, 50^{th}, and 75^{th} percentiles facilitate understanding of score distribution. It is important to note that the inherent kernel density estimation in violin plots may create an illusion of scores surpassing their maximum and minimum limits, despite no actual scores exceeding their threshold.

structured interaction and robot errors. In order to gain further insights into the underlying reasons for this observed correlation, an investigation was conducted to explore the potential influence of system complexity. Therefore, the expected value of "I had to learn more about robots in order to be able to interact with the system" for the structured and unstructured modes is calculated, which is 2.1 ± 1.3, and 2.13 ± 1.25, respectively. This means that even though the subject's expected value on how complex the system is for both structured and unstructured is very close, the subject does not feel safe during the structured tasks when the robot makes errors.

Moreover, in Table II, it is shown that there is a negative correlation between “I am concerned about my privacy when using the robot” and the robot error rate during unstructured tasks and when the second task is unstructured. Some of the participants who started with structured tasks, then followed by unstructured tasks, engaged in testing the system or even trying to make it fail. For example, one of the participants asked the robot: “hey robot, I need corn because my sister wants me to cook pasta with tomato sauce”. The robot was able to detect the intent and fetched corn. Subsequently, the participant asked the robot: “I don’t like eating carrots let’s have some green beans”. The robot was able to understand the intent and fetched green beans. Then the subject asked the robot: “chili is not healthy, but let’s have some”. However, the robot was not able to understand the intent. The participant repeated the request even though the robot was not able to understand the intent, which increased the number of robot errors. In short, some of the participants who completed the unstructured task as their second task felt more comfortable with challenging SousChef.

At last, a negative correlation is found between “It is acceptable for the robot to have much information about the user” and robot errors during the structured tasks. As the robot increasingly made mistakes, the participants became hesitant to share their information with it. It is consistent with the perceived trust in the system “I felt safe using the robot” as shown in Table II. For the rest of the metrics, no significant correlation was found.

From the above-discussed correlations, the hypothesis cannot be unanimously concluded that robot errors adversely affect the overall perception or usability of the system by the user. They might affect some aspects of the perception, but it is on a case-by-case basis; for example, it may vary based on the communication mode of the subject, whether it is structured or unstructured.

Hypothesis 3: To explore the potential impact of past encounters with robot failures during interactions on an individual’s inclination towards a preferred method of instructing the robot, be it structured or unstructured, the Kendall-Tau correlation is used. The Null & Alternate hypothesis and criterion for the statistically significant correlation of the Kendall-Tau method are elucidated in the Hypothesis 2 rationale of the current Section.

Upon analysis, as shown in Table III, it was determined that no statistically significant correlation exists between participant’s experiences in the initial task (Task 1) and their responses in the subsequent task (Task 2) assessing method preference.

Indeed, a lack of correlation was observed between encountered error rate during Task-1 (caused by speech to subject’s intent and robot failures) and any of the other questions in HRCQ as outlined in Table III during Task-2, thereby lending support to Hypothesis 3. This absence of correlation underscores that an individual’s initial method of instruction did not exert a discernible influence on their subsequent preferences in the Task-2 instructional context.

Additionally, Table III shows a negative correlation between the total robot error rate from both tasks and how the subject perceived the predictability of robot actions after completing both tasks regardless of their order. This strongly indicates the unbiasedness of the subjects, as the robot error rate in Task-1 appears to have had no impact on the user feedback metrics for Task-2. However, it was only the combined errors occurring in both Task-1 and Task-2 that influenced the predictability score measured after Task-2.

V. Discussion

As indicated in the results presented in Section IV, it was evident that subjects exhibited a preference for engaging in unstructured, natural spoken language as their primary mode of speech interaction. A noteworthy finding emerged in the manner in which subjects perceived and responded to errors made by the robot. Among the 72 correlations scrutinized in Hypothesis 2, only 4 exhibited any significant correlation.

This particular behavior had been previously identified by [39], a phenomenon commonly referred to as the ”Pratfall Effect” [40]. In accordance with the Pratfall Effect, individuals who are deemed to exhibit a high degree of competence tend to be viewed as more affable when they commit an ordinary error, as opposed to those who lack a similar perception of competence. This phenomenon underscores the nuanced dynamics between perceived competence and the social assessment of likability. Furthermore, Hypothesis 3 did not deviate from this pattern, as errors originating from previous tasks failed to influence how subjects assessed the robot’s usability. This observation aligns with the Pratfall Effect.

Important insights are also given by the open questions as the participants had several suggestions to expand the SousChef capabilities. For example, enabling the robot to fetch multiple objects, or adding additional robot actions (such as opening cabinets), would be beneficial. Some participants were interested in having a more conversational robot, and one participant suggested that it would be good if the robot could understand human emotions. Additionally, several participants would prefer a faster speed for the mobile base.

Moreover, the violin plots in Figure 3 reveal several intriguing insights that can be derived from the data. Specifically, an intriguing pattern emerges with respect to the 75^{th} percentile values within the context of the structured communication mode as opposed to the unstructured mode. In Figure 3-1, the 75^{th} percentile of the unstructured is higher than the structured, which adds credibility to Hypothesis 1, that individuals favored the unstructured way of interaction. In Figure 3-2 for the statement ”I accomplished the given tasks rapidly for structured Vs. unstructured”, both structured and unstructured modes have similar behavior up to the 50^{th} percentile. However, toward the 75^{th} percentile, the subject favored structured methods. One possible reason would be the easiness of commanding the robot since they could just follow the script. For the statement “The robot actions were predictable” (as shown in Figure 3-4), although before the 25^{th} percentile, people perceived the actions of the robot as less predictable in the unstructured compared to the structured,

TABLE II: Hypothesis 2 - Correlation between metric 1 and metric 2. The cells in green show statistically significant correlations.

Metrics		**Kendall-Tau Correlation S= Structured Task, U= Unstructured Task**								
Metric 1 (User Feedback)	**Metric 2 (Robot Error Rate)**	**All Tasks (S+U)**	**S Task**	**U Task**	**Task-1 (S or U)**	**Task-2 (S or U)**	**Task 1 = S Task**	**Task 1 = U Task**	**Task 2 = S Task**	**Task 2 = U Task**
SUS Score	Robot Error Rate	-0.01 (p-value: 0.8726)	-0.01 (p-value: 0.9426)	0.01 (p-value: 0.9424)	0.11 (p-value: 0.4087)	-0.16 (p-value: 0.2483)	0.01 (p-value: 0.9602)	0.24 (p-value: 0.2280)	0.01 (p-value: 0.9599)	-0.26 (p-value: 0.2029)
Robot Actions Were Predictable	Robot Error Rate	-0.11 (p-value: 0.2610)	-0.20 (p-value: 0.1802)	-0.01 (p-value: 0.9545)	-0.03 (p-value: 0.8640)	-0.24 (p-value: 0.1073)	-0.20 (p-value: 0.3448)	0.30 (p-value: 0.1737)	-0.15 (p-value: 0.4818)	-0.28 (p-value: 0.1930)
I Felt Safe Using The Robot	Robot Error Rate	-0.17 (p-value: 0.1275)	-0.34 (p-value: 0.0322)	0.01 (p-value: 0.9411)	-0.09 (p-value: 0.5497)	-0.29 (p-value: 0.0627)	-0.33 (p-value: 0.1491)	0.32 (p-value: 0.1708)	-0.37 (p-value: 0.1018)	-0.21 (p-value: 0.3557)
I Found Robot Easy To Use	Robot Error Rate	-0.18 (p-value: 0.0889)	-0.27 (p-value: 0.0763)	-0.09 (p-value: 0.5513)	-0.13 (p-value: 0.4179)	-0.27 (p-value: 0.0730)	-0.19 (p-value: 0.3840)	0.03 (p-value: 0.8953)	-0.39 (p-value: 0.0822)	-0.16 (p-value: 0.4692)
The Robot Met My Expectations	Robot Error Rate	-0.09 (p-value: 0.3707)	-0.13 (p-value: 0.3743)	-0.05 (p-value: 0.7448)	0.00 (p-value: 1.0000)	-0.22 (p-value: 0.1450)	-0.02 (p-value: 0.9142)	0.05 (p-value: 0.8288)	-0.21 (p-value: 0.3322)	-0.19 (p-value: 0.3928)
I Am Concerned About My Privacy When Using The Robot	Robot Error Rate	-0.04 (p-value: 0.6856)	0.15 (p-value: 0.2859)	-0.33 (p-value: 0.0233)	-0.06 (p-value: 0.7021)	-0.01 (p-value: 0.9550)	-0.03 (p-value: 0.8723)	-0.20 (p-value: 0.3574)	0.31 (p-value: 0.1383)	-0.57 (p-value: 0.0084)
It is Acceptable For The Robot To Have Much Information About The User	Robot Error Rate	-0.03 (p-value: 0.7321)	-0.30 (p-value: 0.0355)	0.22 (p-value: 0.1344)	-0.03 (p-value: 0.8373)	-0.04 (p-value: 0.7598)	-0.26 (p-value: 0.2159)	0.33 (p-value: 0.1263)	-0.30 (p-value: 0.1653)	0.24 (p-value: 0.2617)
I Should Have Full Control Of When and How The Robot Will Assist Me	Robot Error Rate	0.11 (p-value: 0.2778)	0.13 (p-value: 0.3930)	0.08 (p-value: 0.5809)	0.19 (p-value: 0.1917)	-0.01 (p-value: 0.9695)	-0.08 (p-value: 0.7000)	0.29 (p-value: 0.1845)	0.26 (p-value: 0.2197)	-0.37 (p-value: 0.0874)

TABLE III: Hypothesis 3: Correlation between Metric (User Feedback) and Robot error rates Task-1 Robot Error Rate, Task-2 Robot Error Rate, Sum of Robot Error rates (Task-1 Error Rate + Task-2 Error Rate)

Metric (User Feedback)	**Task-1 Robot Error Rate**	**Task-2 Robot Error Rate**	**Sum of Robot Error rates**
SUS Score in Task-2	0.09 (p-value: 0.5160)	-0.16 (p-value: 0.2483)	-0.0 (p-value: 0.9713)
The robot actions were predictable in Task-2	-0.16 (p-value: 0.2802)	-0.24 (p-value: 0.1073)	-0.31 (p-value: 0.0297)
I found the robot easy to use in Task-2	-0.01 (p-value: 0.9513)	-0.27 (p-value: 0.0730)	-0.14 (p-value: 0.3394)
The robot met my expectations in Task-2	-0.13 (p-value: 0.3988)	-0.22 (p-value: 0.1450)	-0.26 (p-value: 0.0752)
I felt safe using the robot in Task-2	0.02 (p-value: 0.8737)	-0.29 (p-value: 0.0627)	-0.15 (p-value: 0.3300)
I am concerned about my privacy when using the robot in Task-2	0.06 (p-value: 0.6793)	-0.01 (p-value: 0.9550)	0.02 (p-value: 0.8956)
It is acceptable for the robot to have much information about the user in Task-2	-0.03 (p-value: 0.8335)	-0.04 (p-value: 0.7598)	-0.11 (p-value: 0.4457)
I should have full control of when and how the robot will assist me in Task-2	0.14 (p-value: 0.3495)	-0.01 (p-value: 0.9695)	0.11 (p-value: 0.4575)

both have the same growth up to the 75^{th} percentile. In Figure 3-8, both modes showed that it is not necessary to learn more about the robot because there is a crest at a low score of around 1 (Statement: I had to learn more about robots in order to be able to interact with the system). However, the unstructured one had a lower 75^{th} percentile, which means that the subject found the unstructured method required less learning. The rest of the violin plots show a similar behavior between structured and unstructured.

In summation, the analysis of violin plots provides a comprehensive perspective on the dynamics between structured and unstructured modes of communication in the context of human-robot interaction. The discernible variations in 75^{th} percentile scores across distinct response categories underscore the nuanced impact of communication paradigms on user perceptions, task execution, and privacy considerations. These insights contribute to a deeper understanding of the interplay between communication modalities and their consequential implications within the realm of HRI research.

VI. Conclusion and Future Work

In this paper, a comparison between structured and unstructured modes of speech communication between a human and a robot is conducted. We collected data from 30 participants during a collaborative cooking task, and SUS and HRCQ survey data were collected during the interaction. This paper found statistically significant evidence that participants preferred the unstructured mode of communication in comparison to the structured one. Additionally, it was proven that there is no significant correlation between the robot error rate and the perceived usability of the robot. Furthermore, the robot error rate in the previous task (Task 1) or the current task (Task 2) has no impact on how the subject perceived robot usability after completing Task 2, which aligns with Pratfall's Effect. Furthermore, the data gathered from the experiments has illuminated significant correlations. These correlations have provided insights into various aspects of human-robot interaction, including safety and privacy concerns. These aspects are thoroughly explored in Sections V and IV.

As our research progresses, we will prioritize the investigation of potential gender-related disparities in the perception and interpretation of robot errors. Furthermore, we intend to expand our efforts by developing a more conversational robot that possesses a deeper understanding of human communication. This direction holds crucial importance for our future endeavors.

Acknowledgment

We express our sincere appreciation to the participants for their priceless contributions to our user study. Additionally, this material is based upon work supported by the National Science Foundation under Grant 2226165.

References

[1] GlobalNewsWire, "Household robots market - growth, trends, covid-19 impact, and forecasts," https://www.globenewswire.com/news-release/2022/03/02/2395266/0/en/Household-Robots-Market-Growth-Trends-COVID-19-Impact-and-Forecasts-2022-2027.html, 2022, [Online; accessed 12-June-2023].

[2] O. Gaggi, G. Quadrio, and A. Bujari, "Accessibility for the visually impaired: State of the art and open issues," *2019 16th IEEE Annual Consumer Communications and Networking Conference, CCNC 2019.*

[3] M. Marge, C. Espy-Wilson, and et al., "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 1 2022.

[4] Q. Xu, Y. Hong, Y. Zhang, W. Chi, and L. Sun, "Grounding language to natural human-robot interaction in robot navigation tasks," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2021, pp. 352–357.

[5] A. Bucker, L. Figueredo, S. Haddadinl, A. Kapoor, S. Ma, and R. Bonatti, "Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 978–984.

[6] M. V. J. Muthugala and A. B. P. Jayasekara, "A review of service robots coping with uncertain information in natural language instructions," *IEEE Access*, vol. 6, pp. 12 913–12 928, 2018.

[7] S. Li and X. Zhang, "Implicit intention communication in human–robot interaction through visual behavior studies," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 437–448, 2017.

[8] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, p. 20, 2023.

[9] K. Kodur, M. Zand, and M. Kyrarini, "Towards robot learning from spoken language," pp. 112–116, 3 2023.

[10] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," *arXiv preprint arXiv:2303.12153*, 2023.

[11] T. X. N. Pham, K. Hayashi, C. Becker-Asano, S. Lacher, and I. Mizuuchi, "Evaluating the usability and users' acceptance of a kitchen assistant robot in household environment," in *2017 26th IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 987–992.

[12] B. Cagiltay, H.-R. Ho, J. E. Michaelis, and B. Mutlu, "Investigating family perceptions and design preferences for an in-home robot," in *Proceedings of the Interaction Design & Children Conf.*, 2020.

[13] M. Yamamoto, Y. Hu, E. Coronado, and G. Venture, "Impression evaluation of robot's behavior when assisting human in a cooking task," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 743–748.

[14] W. Y. G. Louie and G. Nejat, "A social robot learning to facilitate an assistive group-based activity from non-expert caregivers," *Int. J. of Social Robotics*, vol. 12, pp. 1159–1176, 11 2020.

[15] A. Angleraud, A. M. Sefat, M. Netzev, and R. Pieters, "Coordinating shared tasks in human-robot collaboration by commands," *Frontiers in Robotics and AI*, vol. 8, p. 734548, 10 2021.

[16] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction," *Sensors*, 2022.

[17] H. Chen, M. C. Leu, and Z. Yin, "Real-time multi-modal human-robot collaboration using gestures and speech," *J. of Manufacturing Science and Engineering*, vol. 144, 10 2022.

[18] I. Giorgi, A. Cangelosi, and G. L. Masala, "Learning actions from natural language instructions using an on-world embodied cognitive architecture," *Frontiers in Neurorobotics*, vol. 15, p. 48, 5 2021.

[19] M. A. et al, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022.

[20] Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with chatgpt," *IEEE Access*, vol. 11, pp. 55 748–55 754, 2023.

[21] M. Kyrarini, Q. Zheng, M. A. Haseeb, and A. Gräser, "Robot learning of assistive manipulation tasks by demonstration via head gesture-based interface," in *2019 IEEE 16th Int. Conf. on Rehabilitation Robotics (ICORR)*. IEEE, 2019, pp. 1139–1146.

[22] A. Chowdhery and et al, "Palm: Scaling language modeling with pathways," 2022.

[23] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," 2017. [Online]. Available: https://arxiv.org/abs/1706.04261

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] T. Wolf, L. Debut, and et al, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020.

[29] S. Tiwari, "An introduction to qr code technology," in *2016 international conference on information technology (ICIT)*. IEEE, 2016, pp. 39–44.

[30] V. Jain, Y. Jain, H. Dhingra, D. Saini, M. Taplamacioglu, and M. Saka, "A systematic literature review on qr code detection and pre-processing," *Int. J. on Technical and Physical Problems of Engineering*, 2021.

[31] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage." 1990.

[32] J. Brooke, "Sus: a "quick and dirty'usability," *Usability evaluation in industry*, vol. 189, no. 3, pp. 189–194, 1996.

[33] J. Schmidtler, K. Bengler, F. Dimeas, and A. Campeau-Lecours, "A questionnaire for the evaluation of physical assistive devices (quead): Testing usability and acceptance in physical human-robot interaction," in *2017 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 2017.

[34] G. Charalambous, S. Fletcher, and P. Webb, "The development of a scale to evaluate trust in industrial human-robot collaboration," *International Journal of Social Robotics*, vol. 8, pp. 193–209, 2016.

[35] Student, "THE PROBABLE ERROR OF A MEAN," *Biometrika*, vol. 6, no. 1, pp. 1–25, 03 1908. [Online]. Available: https://doi.org/10.1093/biomet/6.1.1

[36] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 12 1965.

[37] H. Abdi, "The kendall rank correlation coefficient," 2007. [Online]. Available: http://www.utd.edu/

[38] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia and Analgesia*, vol. 126, pp. 1763–1768, 5 2018.

[39] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To err is robot: How humans assess and act toward an erroneous social robot," *Frontiers Robotics AI*, vol. 4, p. 251625, 5 2017.

[40] E. Aronson, B. Willerman, and J. Floyd, "The effect of a pratfall on increasing interpersonal attractiveness," *Psychonomic Science 1966 4:6*, vol. 4, pp. 227–228, 2 2014.