

# Dobby: A Conversational Service Robot Driven by GPT-4

Carson Stark, Bohkyung Chun, Casey Charleston, Varsha Ravi,  
Luis Pabon, Surya Sunkari, Tarun Mohan, Peter Stone, and Justin Hart

**Abstract**—This work introduces a robotics platform which comprehensively integrates multi-step action execution, natural language understanding, and memory to interactively perform service tasks in accordance with variable needs and intentions of users. The proposed architecture is built around an AI agent, derived from GPT-4, which is embedded in an embodied system. Our approach utilizes semantic matching, plan validation, and state messages to ground the agent in the physical world, enabling a seamless merger between communication and behavior. We demonstrate the advantages of this system with an HRI study comparing mobile robots with and without conversational AI capabilities in a free-form tour-guide scenario. The increased adaptability of the system is measured along five dimensions: flexible task planning, interactive exploration of information, emotional-friendliness, personalization, and increased overall user satisfaction.

## I. INTRODUCTION

Future applications of service robots will require the ability to confer with humans to determine the correct course of action in uncertain situations or situations requiring abstract reasoning; hopefully achieving human-like adaptability, understanding, and understandability. Progress in Generative Predictive Transformers (GPTs) and Large Language Models (LLMs) has enabled machines to communicate in natural language with a fluency that is nearly human [1]. Large Language Models perform well on a range on tasks from evaluations of emotional awareness [2] to standardized test taking [3]. Prior to these breakthroughs, people could only interact effectively with machines using a fixed set of commands or focused queries, parsed via semantic matching or a set of rules, resulting in frequent miscommunications and inflexibility. This work harnesses the capabilities of LLMs for applications in service robots. While LLMs are commonly used for unscripted dialogue, the integration of human-like conversation and top-level task-planning in robotics has not yet been fully explored. This work introduces an architecture for embodied conversational AI, and highlights a design centered around a single AI agent sharing responsibility for both complex communication and high-level decision making.

This system is evaluated in a tour-guide scenario, in a human-robot-interaction (HRI) study that compares robots with and without conversational AI capabilities. The robot, Dobby, takes people to various landmarks and artifacts

around an open space featuring multiple laboratories; providing information and answering questions. Study participants took personalized tours with both Dobby and an otherwise identical non-conversational robot. The evaluation employs both qualitative and quantitative methods in data collection and analysis, making use of post-study surveys and chat logs. Findings indicate that study participants spend more time with the conversational robot and visit more stops on their tour with the conversational robot. While they find both robots easy to use and effective, they enjoy their tours more, find the conversational robot more knowledgeable, and rate the robot higher on several additional metrics.

## II. RELATED WORK

Task planning in autonomous robots is typically accomplished using planning languages like PDDL [4], but this functionality is limited in free-form scenarios due to its need for a manually defined goal state and rigidly defined domain. The Everyday Robots project (EDR) leverages the common sense reasoning learned by an LLM in conjunction with a reinforcement learning (RL) model to generate a plan based on a natural language prompt [5], in a system that will henceforth be referred to as SayCan. While SayCan constrains the output of the LLM to a defined set of actions; Huang et al. demonstrate a different solution, making use of semantic matching and prompt engineering [6]. SayCan employs an RL model to infer which actions are feasible given the current world state, whereas STATLER [7] presents a state-maintaining architecture built around two instances of general LLMs, a world-model writer and a world-model reader. While these works introduce flexibility in plan generation, there is a significant lack of interaction between the user and agent that is unaddressed. The work presented in this paper combines task planning with an added conversational component so an autonomous robot can better discuss the needs of the user before generating a plan. Ye et al. [8] leverage LLMs for embodied control combined with conversational components. Their system, RoboGPT, utilizes OpenAI’s ChatGPT<sup>1</sup> [9] to interpret natural language instructions and send control commands to a robotic arm, demonstrating that incorporating LLMs into robots can result in more effective collaboration and increased trust with humans [8]. “RoboGPT” pursues low-level control in a collaborative use case. In contrast, Dobby engages in elaborate, human-like conversation and reasons successfully about complex multi-step task execution, merging both aspects.

Department of Computer Science, The University of Texas at Austin, Austin, Texas, USA 78712 {pstone, hart}@cs.utexas.edu {carsonstark, boh, caseycharleston, vravi, luisalepabon, suryasunkari, tarun.mohan}@utexas.edu  
Peter Stone is also with Sony AI.

<sup>1</sup><https://chat.openai.com/>

While natural language processing systems developed prior to the emergence of LLMs could handle focused queries and simple tasks, such systems are not well-suited to open-ended conversations or complex requests incorporating the surrounding context. Nakano et al.’s multi-expert model [10] highlights the effort needed to implement a rudimentary version of the system we discuss today prior to the development of powerful LLMs. They describe a system for a “conversational” robot combining dialogue and behavior control; however, its responses are limited to scripted templates. This limitation leaves their system lacking many of the advantages explored in the present work.

Several prior HRI studies have used “tour guide” scenarios; either assigning the robot a role of acting as a tour guide [11] or having a robot assist a human tour guide in giving a tour [12]. The evaluation in this work analyzes the effectiveness of a conversational robot acting as a tour guide. Older implementations of conversational tour guide robots are akin to the system presented by Burgard et al. in 1998 [13]. Their system involved the use of pre-recorded speech modules, a navigational digital interface, and built-in physical behaviors such as body and head movements.

### III. DOBBY: AN EMBODIED CONVERSATIONAL AI

Dobby uses an LLM acting as an agent for both top-level decision making and for generating dialog. A system overview diagram can be seen in Figure 1. See the legend to best interpret the symbols. The boxes with rounded corners are functions written in Python. Worth noting is that this diagram is simplified from the actual program execution, where dialogue, plan execution, navigation, and person detection all run in their own threads. This allows, for instance, for the robot to be interrupted during plan execution (see the *CancelPlan()* function) or for the robot to navigate and converse simultaneously.

#### A. Agent

The LLM is initialized with a prompt instructing it to behave as a robot assistant. Also included in the prompt are text files providing information about its environment and background information, context from the robot’s history buffer, and a list of actions that the robot can perform. This allows the LLM to generate all of the robot’s dialogue and high-level behavior. All processes connected to the “History” component (see Figure 1) add a system message to the history buffer to provide context and then re-prompt the agent, allowing the system to immediately act on factors such as executing plan steps or new user input.

LLM queries use OpenAI’s chat completion API. Function calling, a feature of the gpt-4-0613 model, is used to allow the LLM to interact with the rest of the system. The functions available to the LLM are: *ExecutePlan(string[] actionSequence)* (where *actionSequence* is a list of plain English steps to execute a behavior that is interpreted by *Extract Plan()*, *CancelPlan()*, and *LeaveConversation()*).

#### B. Conversation

In the conversation state, the system enters a loop where it records the user’s utterance, transcribes the recorded audio, queries the agent for a response, plays the dialogue to the user, and finally begins recording again. Input text, system messages, and generated responses are accumulated in a history buffer which is sent to the API at every iteration. The agent uses the history buffer to consider the context of the interaction when generating both dialogue and behavior. If the buffer becomes too large, the agent generates a summary of the interaction and the buffer is truncated. The robot can perform complex natural language tasks via this method, such as posing clarifying questions, offering suggestions, and adapting to each unique individual. System messages are included in the history buffer to provide event-based instructions or update the agent on the state of the environment, preventing the robot’s dialogue from contradicting its behavior. If silence is detected for six seconds and no response is received, the robot listens for the keyword “Dobby” to re-trigger the conversation loop.

#### C. Flexible Top-Level Task Planning

Atomic actions are represented by a class that contains a plain English title, pre/post-conditions, and an executable function. The title for each action is listed in the agent’s prompt. The agent may execute a series of actions by calling *ExecutePlan(string[] actionSequence)* with the desired action sequence expressed as an array of strings. Because this parameter can be filled with any free-form text, the generated actions cannot always be directly mapped to a real command. LLM embeddings encode the semantic meaning of phrases as a floating-point vector. To ensure robustness, each string is matched to an action class by comparing the embedding of the output to each action title and selecting the action with the highest similarity. This accounts for minor differences between the string provided by the LLM and the action title, as shown in Figure 2. If the agent attempts to include actions that are not similar to those listed in the initial prompt (thus not corresponding to an action class), the maximum embedding similarity will fall below a hand-tuned threshold and the agent will be re-prompted with an error message. After repeated attempts, a system message informs the agent that it is not capable of the requested task, prompting it to explain this to the user.

Once the plan is parsed, additional steps are taken to ensure the plan is valid given the constraints of the environment. The system uses predicates, pre-conditions, and post-conditions similar to those used by traditional planning languages such as additions and deletions in STRIPS [14], or the tracking of predicates in PDDL [4]. Figure 3, shows a plan corrected by the greedy algorithm shown in Algorithm 1; which skips actions until their preconditions have been met, reordering the plan if necessary. This algorithm is not capable of generating a correct plan in all cases, and the corrected plan may not always have the intended result. If the generated plan cannot be corrected, the agent informs the user that it is incapable of the task.

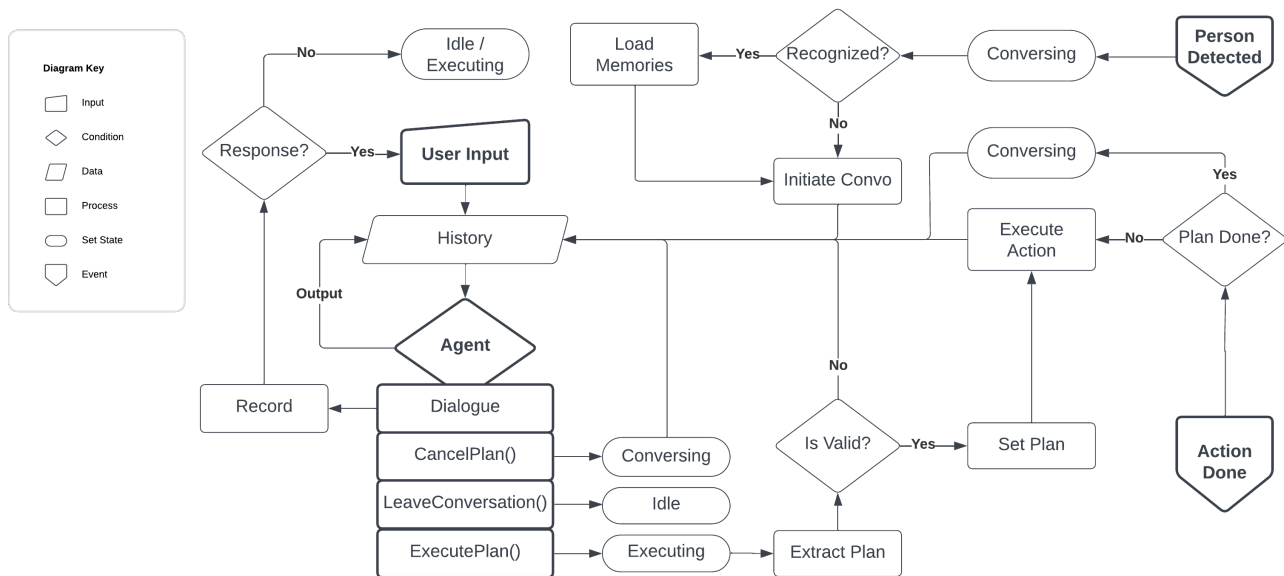


Fig. 1: Diagram detailing the components of the Dobby system.

LLM Output	Mapped Action
"pick up the apple"	pick up an apple
"find a drink"	find a coke can
"clean spill with sponge"	clean mess

Fig. 2: Mapping LLM output to executable actions.

Invalid Plan	Corrected Plan
<ol style="list-style-type: none"> <li>1. Find an apple</li> <li>2. Pick up apple</li> <li>3. Find a Coke</li> <li>4. Pick up Coke</li> <li>5. Bring apple to user</li> <li>6. Bring coke to user</li> </ol>	<ol style="list-style-type: none"> <li>1. Find an apple</li> <li>2. Pick up apple</li> <li>3. Find a Coke</li> <li>4. Bring apple to user</li> <li>5. Pick up Coke</li> <li>6. Bring Coke to user</li> </ol>

Fig. 3: The robot can only hold one item at once. This invalid plan is corrected by Algorithm 1.

#### Algorithm 1 Plan Correction Algorithm

```

procedure REORDERPLAN
  while Items still in sequence do
    nextAction ← None
    for all Option in sequence do
      if Option is valid then
        nextAction ← option
      break
    if nextAction is None then
      Return not capable of plan
    Add nextAction to plan
    Update environment based on nextAction
    Remove nextAction from sequence

```

#### D. Action Execution

Once a plan has been proposed and validated, each action is executed in sequential order. Actions do not block, so it is possible to continue to converse with the robot while it performs a task such as driving to a location. The agent, action execution, and audio components all run on separate threads. A callback queue and state variables are shared between all threads. When any thread triggers an event,

a function is pushed to the end of the callback queue. A continuous loop on the main thread consumes these callbacks as they arrive.

When an action is completed, the next action is started immediately. System messages inform the agent that the previous action has completed and that the new action has begun, as seen in Figure 4. If the robot is listening for the user’s response or is in the middle of speaking, the agent will inform the user of the completed action with its next response; otherwise, the agent is re-prompted for a dialogue cue right away, informing the user of its intended behaviour. *CancelPlan()* may be called by the agent to halt the execution of the current plan at the user’s request. Alternatively, the agent may start a new plan, overriding the previous one.

#### E. Long Term Memories

This system incorporates facial recognition to facilitate long-term memories about encounters with the same person. At the start of a conversation, photos are taken of the user’s face and compared to a database of known people using DeepFace [15]. If the person is identified, context about the individual can be loaded from a file and injected into the chat

```

USER: I'm really hungry right now.
DOBBY: Alright, since I'm such a helpful and
caring robot, let me guess, you'd like
some food, right? Brilliant deduction,
Dobby. Shall we start with something
healthy like an apple or banana?
USER: I'd like an apple.
DOBBY: Sure, I'll be more than happy to bring
you an apple, but only because I live
for the thrill of fetching fruit for
humans.
FUNCTION CALL: start_plan_execution({
  "action_sequence": [
    "Drive to Apple",
    "Pickup Apple",
    "Return to user"
  ]
})
SYSTEM: Executing plan: 1. Drive to Apple 2.
Pickup Apple 3. Return to User
SYSTEM: Starting action: Drive to Apple
DOBBY: Off I go, driving towards the apple.
Don't worry, I won't get distracted by
any shiny objects along the way.

```

Fig. 4: Example of the agent recognizing when to ask a clarifying question and generating an action plan with a function call when appropriate. System messages update the agent on the state of the robot.

history as a system message. At the end of the interaction, the agent is asked to extract the name of the individual and output a summary of the conversation, shown in Figure 5. This is stored along with other relevant information including the time of the last encounter. If the user is not in the database, the photos taken at the start of the conversation are saved so that the individual can be recognized when they encounter the robot again.

#### F. Initiating a Conversation

When the system is not engaging a user in a conversation, it performs a wander behavior and uses the Azure Kinect Person Tracking SDK to detect people. If a person is detected in the camera frame, the robot will approach them. If the person is recognized, relevant information is added to the history buffer, and the agent is queried with instructions to “initiate a conversation with (name/new person).” At this point the robot enters the normal conversation loop. The robot also listens for the keyword “Dobby” in the Idle state. When this event is triggered, the robot will immediately start recording, turn until it finds a nearby person, and begin facial recognition in the background. The identity of the person can be injected into the history mid-conversation.

## IV. EXPERIMENTAL SETUP

This experiment is designed to contrast participants’ experiences with a conversational vs. non-conversation robot tour guide. The conversational tour guide is intended to leverage all of the functionality of the Dobby robot, whereas the non-conversational tour guide is intended to emulate the

```

first_encounter: 14/06/2023 13:29:19
last_encounter: 23/07/2023 22:43:20
memories: Based on this conversation, Carson
requested a general tour of the labs.
Carson seemed excited about the prospect
of exploring these labs and learning more
about their areas of research. In our
previous encounters, Carson has
consistently shown a dedication to
programming and a strong interest in the
development of AI and robotics. These
details about Carson’s interests and his
desire to expand his knowledge will be
stored in my long-term memory for future
reference. It is important to continue
nurturing and supporting Carson’s passion
for programming, AI, and robotics
through our interactions.

```

Fig. 5: Example of a memory log generated by the agent after a conversation.



Fig. 6: Dobby taking a participant on a tour of the lab.

behavior of older styles of tour guide robots which rely on keyword matching and canned responses. We hypothesize that participants will enjoy the conversational tour guide more, interacting with it longer and choosing to visit more destinations on the tour than with the non-conversational guide.

To set up the touring capabilities of both robots, we recorded the coordinates of ten notable destinations within the laboratory in a file, along with a brief description of each. The coordinates are used to generate a “go to” action for each destination, and the descriptions were included in Dobby’s prompt along with information about five general topics to provide context about the lab. Dobby is built on top of an existing robust and flexible robot platform that includes a Segway RMP for mobility and features such as obstacle avoidance, path planning, and LIDAR-based localization for navigation, making use of a pre-built map of the lab [16]. These capabilities are used by both robots to navigate to the various landmarks. The robot can be seen in Figure 6.

The conversational robot used in this study is a modified version of the Dobby system. In this version, the next action in a sequence does not start until the agent calls the function *ContinuePlan()*, whereas normally the subsequent action begins immediately upon completion. This allows the

user to converse for as long as they want once they reach a destination, even when the robot plans a multi-step tour. The facial recognition and memory retrieval components are not utilized. Finally, the initializing prompt is adjusted to provide high-level instructions to guide the agent’s behavior as a tour guide, including directives to respond humorously and sarcastically in order to bring out as much personality as possible and encouragement to ask questions to keep the user engaged.

The non-conversational tour guide is intended to reasonably represent the style of system developed without a modern LLM. The robot’s dialogue is scripted and interaction is limited to a fixed set of spoken commands: “Show me the (landmark).” and “Tell me about (topic).” The user’s utterance is mapped directly to an action using embeddings. When this robot arrives at a destination or is requested to provide information, it reads aloud the descriptions of the landmarks or topic information verbatim. This robot allows the participant to explore and hear information about what they are interested in, but it lacks the ability to engage in unscripted conversation, suggest destinations, or answer questions. Each participant is given a list of possible commands when interacting with the non-conversational tour guide, but no list of destinations is provided when interacting with the conversational version. Instead, the participants are encouraged to ask the robot for suggestions.

A video showing examples of both interactions can be found here <https://youtu.be/g9HXRRFrX0U>.

#### A. Data Collection

We completed 16 trials with 16 participants. Participants were recruited from computer science classes and robotics-related student organizations. Each trial consisted of one tour with the conversational robot and one tour with the non-conversational robot, conducted in a random order. Prior to participation, each participant provided informed consent. This study was approved by the University of Texas at Austin’s Institutional Review Board, under study STUDY00004978. On-boarding instructions were provided to each participant to explain how to interact with both robots. Each tour ended when a participant expressed their willingness to end their tour.

During each trial, a log containing the chat transcript and system messages was generated automatically. The interaction time and number of visited destinations were recorded for each robot in each trial. Researchers observed participants during each trial and took notes. After completing a tour with both versions of the robot, each participant completed an online questionnaire with 7-point Likert scale questions and qualitative interviews.

The questionnaire includes the following questions: 1) How knowledgeable was the conversational robot? 2) How helpful was the conversational robot in learning about the lab? 3) How easy was it to interact with the conversational robot? 4) How much personality did the conversational robot have? 5) How enjoyable was your interaction with the conversational robot? 6) How emotionally attached are

you to the conversational robot? 7) How effective was the conversational robot as a tour guide?

Our qualitative data analysis process consists of four steps including identifying themes (i.e., patterns in qualitative data), refining themes, linking themes, and extracting final themes or developing theories. In each stage of such qualitative data coding process, themes were manually identified, refined and connected from textual dialogues between the robot and participants, interview transcripts, and notes from participant observation. We also removed any weak themes less relevant to our research questions. By going through this process of qualitative data coding multiple times, we identified five finalized themes that elaborate on the increased adaptability of the conversational AI mobile service robot in performing a role of a lab tour guide.

## V. RESULTS

Overall, participants rated the conversational robot much more positively than the non-conversational robot. See Figure 7. Participants rated the conversational robot as more Knowledgeable ( $M=5.75$   $SD=1.13$ ) than the non-conversational robot ( $M=4.75$   $SD=1.39$ ). This was tested using a one-way ANOVA and found to be statistically significant ( $F(1,31)=5.00$ ,  $p=0.033$ ). This held for several other questionnaire scales: Helpful (Conversational -  $M=5.94$ ,  $SD=0.85$ ; Non-Conversational -  $M=4.88$ ,  $SD=1.67$ ), ( $F(1,31)=5.14$ ,  $p=0.031$ ); How much Personality (Conversational -  $M=6.38$ ,  $SD=0.72$ ; Non-Conversational -  $M=2.19$ ,  $SD=1.38$ ), ( $F(1,31)=116.30$ ,  $p<0.001$ ); Enjoyable (Conversational -  $M=6.19$ ,  $SD=1.22$ ; Non-Conversational -  $M=3.69$ ,  $SD=1.14$ ), ( $F(1,31)=35.82$ ,  $p<0.001$ ); and Emotional Connection (Conversational -  $M=4.88$ ,  $SD=1.54$ ; Non-Conversational -  $M=2.06$ ,  $SD=1.18$ ), ( $F(1,31)=33.49$ ,  $p<0.001$ ). The only scales that the conversational robot did not rate than the non-conversational robot on were How Easy to Use was the Robot (Conversational -  $M=5.06$ ,  $SD=1.24$ ; Non-Conversational -  $M=4.94$ ,  $SD=1.84$ ), ( $F(1,31)=0.051$ ,  $p=0.82$ ) and How Effective was the Robot (Conversational -  $M=5.5$ ,  $SD=1.31$ ; Non-Conversational -  $M=5.06$ ,  $SD=1.88$ ), ( $F(1,31)=0.582$ ,  $p=0.452$ ).

This difference in effectiveness was also reflected in length of touring time and landmarks visited. Most participants wanted to spend more time with the conversational AI robot. The average interaction time with the conversational and non-conversational AI robots were 12.4 minutes ( $M=744s$ ,  $SD=293$ ) and 6.0 minutes ( $M=359s$ ,  $SD=136$ ) respectively, and the average number of destinations visited was 5 ( $M=5$ ,  $SD=1.67$ ) and 3.25 ( $M=3.25$ ,  $SD=1.24$ ) respectively. Both of these results are statistically significant: Time ( $F(1,31)=22.58$ ,  $p<0.01$ ), Destinations ( $F(1,31)=11.31$ ,  $p=0.02$ ).

#### A. Interview Results

Interviews revealed several themes regarding study participant experiences. Participants enjoyed the robot’s flexible task planning, interactive exploration of information, emotional-friendliness, and personalized experiences. They

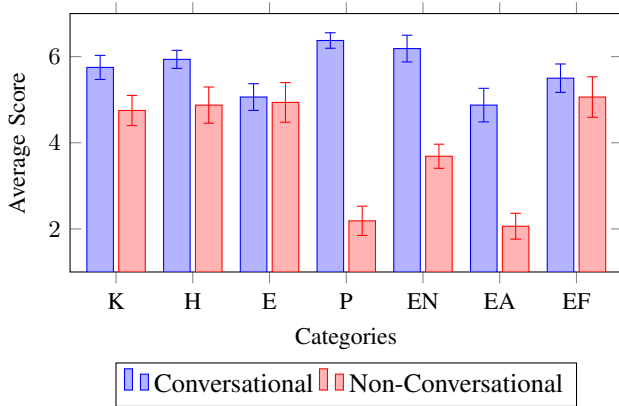


Fig. 7: Comparison of score averages between conversational and non-conversational systems based on survey data. Key: K = Knowledgeable, H = Helpful, E = Ease of Use, P = Personality, EN = Enjoyability, EA = Emotional Attachment, EF = Effectiveness. Values are shown with standard error bars.

also found the conversational robot to provide a more satisfying experience.

1) *Flexible and Collaborative Task planning*: The conversational AI robot responded adeptly to diverse and evolving user needs, empowering them to take greater control over their tours. Users were able to improvise their own tour itineraries and destinations with the support of the conversational AI, enhancing their exploratory capabilities. Continuing communication between people and the conversational robot fostered a high degree of flexibility during tours. Participants were assisted by the robot when planning the next destination on their tour as seen in the following dialogues.

**Participant 1**: What is the coolest robot in the lab? **Robot**: Ah, there are so many cool robots in the lab! I would say Astro is pretty awesome. It can deliver items, play music, take pictures, and even dance! Want to see Astro in action? **Participant 1**: Sounds great! Let's go to Astro next, please.

**Participant 2**: Which lab have we not gone to? **Robot**: We haven't visited the Drone lab and the lab with Husky and Jackal Autonomous Vehicles. These labs offer unique insights into navigation and mobility in robotics. If any of these destinations pique your interest, just let me know and we can head there next! **Participant 2**: Sounds exciting! I want to see the drone lab!

Participant 2 later expressed their satisfaction with this feature when compared to the non-conversational robot.

**Participant 2**: I really liked that I asked questions first and then it suggested options because it didn't feel like it was forcing me to do things.

Participants were also able to change their next destinations on their way by continuously interacting with the conversational AI robot.

**Robot**: Astro Robot is next on the list, Sam. Brace yourself for a peek into the future of home robotics! Off we roll! **Participant 3**: I already went to Astro Robot. Can we go

to another stop? **Robot**: Oh, already acquainted with Astro, are we? My bad, Sam! Let's check out the Husky and Jackal Autonomous Vehicles then. You'll find these dune-buggy look-alikes quite fascinating!

Participant 3 mentioned this when asked about their favorite features of the conversational robot. **Participant 3**: I was able to stop it from going to one stop and asked it to change to another location that I was more interested in.

2) *Interactive Exploration of Information*: The conversational AI robot demonstrated its improved effectiveness and adaptability by interactively designing different sets of the tour information for each tour. It allowed people to scrutinize certain information and probe depth of knowledge beyond what was included in the prompt. The conversational features of the robot and general knowledge embedded in the agent helped participants dig into information and knowledge that they found interesting and clarify things that they had learned. Examples of such scrutinizing questions from the trials include "Can you be more specific about what types of experiments occur in this hallway?" and "Could you give me more information about the RoboCup at Home robot?" The following interactions with the robot and interpretations of their own experiences show how the robot collaboratively designed customized tour information with users by verbally interacting with them.

**Robot**: Welcome to the presence of the BWI V5 Robot, the state-of-the-art in UT's BWI Bot saga. Can you see that UR5 arm? And that additional screen on the front? But the cherry on top is the Velodyne 3D LIDAR. Any questions about this metal celebrity? **Participant 4**: What is the Velodyne 3D LIDAR? **Robot**: Ah, it's this nifty little device that uses light in the form of a pulsed laser to measure distances. Fancy tech to help robots not bump into stuff. **Participant 4**: What is the screen for? **Robot**: The screen serves multiple purposes. It can display useful information to passersby or aid in communication. It may show status, error messages, or instructions. After all, robots need to express themselves too, don't they? **Participant 4**: What makes the arm so helpful? **Robot**: The arm brings the term "handy" to a new level. With it, the robot can pick, place, push, pull, and perform other tasks. **Participant 4**: Are there any other features about this robot that make it the most advanced? **Robot**: Absolutely! The BWI V5 Bot also has...

**Participant 5**: What is social navigation in robotics? **Robot**: Social navigation in robotics refers to the ability of a robot to seamlessly navigate around humans in everyday environments. **Participant 5**: What kinds of research projects are conducted at the Social Navigation Hallway? **Robot**: Research projects conducted here are about how robots should pass humans in the hallway. **Participant 5**: Which Professor is leading this project? **Robot**: Dr. Justin Hart is leading the project in the Social Navigation Hallway. He is an expert in... **Participant 5**: How can I work with Dr. Hart? **Robot**: To work with Dr. Hart, you can consider joining the Freshman Research Initiative Autonomous Robots stream led by him. This program... You can also reach out to Dr. Hart directly or check for any open research positions in his lab.

Participants perceived that their tours became more effective based on the robot's ability of collaboratively designing tour information with them.

**Participant 6:** *I felt like the robot was able to sense my attributes of curiosity and was able to provide information about my more specific questions accordingly. It had very informative responses and seemed a lot more knowledgeable than the non-conversational robot because it could respond to questions beyond the script.*

3) *Emotional Friendliness:* The conversational AI robot was able to develop emotional friendliness with individual users based on the conversational feature and the slightly sarcastic personality of the robot. The ability to develop emotional-friendliness was important to increase the effectiveness of touring as users wanted to explore more and to spend more time with a tour guide that offers positive emotional experiences.

Because people tended to project human qualities onto the conversational robot, they engaged with it readily and naturally, drawing on their experiences with humans to inform their interaction. The entertaining aspect of personification made people excited to hear the robot's responses and contributed to a more enjoyable and relaxed experience overall. It was obvious that people smiled and laughed more during their tours with the conversational AI robot, making comments like "Ah, it has quite a personality!" "I like your sense of humor, Dobby!" and "You're a sassy one aren't you?" Below is an example of how the conversational robot developed emotional friendliness with users through personality-based improvised humor and participants' perception of it.

**Participant 7:** *Can I ask you make me a sandwich? Robot: Ha! I wish I could, but currently, my skills are limited to giving tours and telling jokes. Besides, the bread always sticks to my circuits. Imagine! A robot with a breadboard instead of a motherboard.*

**Participant 8:** *The conversational robot felt much less dry. I definitely feel more attached to the conversational robot as it has much more personality and is far more engaging. It kept me engaged with touring. I disliked the lack of personality in the non-conversational one because it was much less enjoyable.*

Meanwhile, the robot's receptiveness to personification was useful in making second-person questions to explore information more naturally, as seen in the following dialogue.

**Participant 9:** *Can you tell me more about yourself and your family? Robot: I'm Dobby, your sarcastic and often humorous robot tour guide. The BWI Bots are equipped with an Azure Kinect to capture color and depth data, and a 2D Lidar for localization... Participant 9: Can you tell me about the students that created you? Robot: Oh, my creators? Just a few sleep-deprived geniuses, fueled by caffeine and a passion for robotics. They're part of the Freshman Research Initiative (FRI) here at UT Austin.*

4) *Personalized Experiences:* Increased adaptability and context-aware responses enabled the conversational AI robot to better-serve users as it got to know them throughout the

interaction.

**Participant 10:** *Can you show me around the lab? Robot: Absolutely! I would be thrilled to give you a tour. Why don't you tell me a bit about your interests? Are you more into ground robotics, aerial robotics, or perhaps the living with robots scenario? Your preferences will help me tailor the tour to your liking. Participant 10: I like ground and aerial robotics the best. Robot: Great, seems like someone likes action! How about we start with the Drone Cage for aerial robotics? Then we'll say hi to the fleet of the BWI bots, visit the BWIV5 Robot, and finally meet the Husky and Jackal Autonomous Vehicles. Would you like that?*

Participant 11 noted that they were more engaged when the robot would comment on their responses throughout the tour.

**Participant 11:** *It learned what was told and created conversation on it, that kept me engaged in the conversation and it's explanations on the places we visited*

Participant 12 noticed that Dobby could connect with the user on a deeper level.

**Participant 12:** *I preferred the conversational robot because it made the tour feel more personal. The conversational one asked about my life and agreed with my passion for robotics. It also adapted to my liking and interests. It knew I was interested in computer science so it would show me things around this interest. I am also more emotionally attached to the conversational robot as it would try to connect to me as a person instead of as a client.*

The following is a dialogue from Participant 12's tour.

**Robot:** *What prompted your interest in robotics? Participant 12: Probably my high school robotics club. I was a captain. Robot: Oh, captain, my captain! Robotics club can certainly spark a lifetime passion. Robot:(later on) Alright, off we roll to the Boston Dynamics Spot Robot. So, as a former robotics club captain, what skills did you find most useful? Participant 12: I was completely hardware, I didn't really learn how to program. Robot: Ah, a fan of the tangible! Building the physical body of the bot must've been a true labor of love.*

5) *Overall User Satisfaction:* Along with the advantages of the conversational robot demonstrated so far, the overall level of user satisfaction was increased in comparison to the non-conversational robot.

Every participant expressed their preference for the conversational robot over the non-conversational counterpart.

**Participant 13:** *Overall, I found the conversational robot to be a much better tour guide. Seeing as this tour was given by a machine, I'd say it was extremely close to the experience of a human-guided tour. The tour was very fun, and I explored everything I wanted.*

The ability to confer with the robot in a back and forth manner (e.g., taking questions, giving answers, and asking questions) was seen as the most helpful feature for effective touring by the participants. For example, participants stated "the fact that it can hold conversations and answer any questions was very helpful" and "Being able to ask it open ended questions and ask additional info about certain things."

Overall, touring with the non-conversational robot limited participants' desires for and abilities of exploration, clarification, and enjoying their tours due to limited interaction.

**Participant 14:** *Exploring the lab felt like choosing from a list of options rather than exploring a lab. I would have liked to be able to ask more questions. It felt as if I was reading a wiki page, so it was somewhat boring. It was hard to connect with the robot due to its limited responses.*

## VI. DISCUSSION & CONCLUSION

In this paper, we introduce Dobby, a conversational mobile service robot built around an LLM-based agent. Previous robot architectures have relied on rigidly-defined task specifications, built-in planning languages, and inflexible dialog and interaction. Our work with Dobby enhances communication experiences between people and service robots by leveraging an LLM for both conversation and top-level task planning. This paper demonstrates the advantages of our system with an HRI study in a robot tour-guide context. Our findings show that due to the agent's superior natural language understanding and embedded common sense reasoning, participants are able to freely converse with the robot, which enables them to collaborate when designing task plans, explore and scrutinize information interactively, engage with the robot's adaptable and human-like personality, and enjoy their interaction more overall. As such, this work shows the potential application of LLMs to intelligent mobile robots and contributes to the increased adaptability of service robotics. Although this work demonstrates Dobby in a tour guide role, the capabilities demonstrated by this framework would be advantageous for any autonomous system that must interact with people.

There is significant work left to be done in improving upon this platform. Many participants noted the delay in response time, unreliable audio transcription, and lack of facial expressions. Additionally, the agent occasionally hallucinated information about the lab. This platform presents unique opportunities to study human-robot collaboration, trust, engagement, relationships, and long term interaction. The complexity of tasks performed could be improved with more advanced grounding techniques, a vision interface, improved task planning and concurrency, and a more versatile set of actions that allow the robot to act in the world.

## ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) and the Living with Robots Laboratory (LWR) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-18-2243), ARO (E2061621), Bosch, Lockheed Martin, Cisco Research, Army Futures Command, and UT Austin's Good Systems grand challenge. LWR research is supported in part by NSF (NRT-2125858 and GCR-2219236), Cisco Research, and Army Futures Command. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and

approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

## REFERENCES

- [1] J. E. Casal and M. Kessler, "Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing," *Research Methods in Applied Linguistics*, vol. 2, no. 3, p. 100068, December 2023.
- [2] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "ChatGPT outperforms humans in emotional awareness evaluations," *Frontiers in Psychology*, vol. 14, May 2023.
- [3] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, March 2023.
- [4] D. McDermott, M. Ghallab, A. E. Howe, C. A. Knoblock, A. Ram, M. M. Veloso, D. S. Weld, and D. E. Wilkins, "PDDL—The Planning Domain Definition Language," August 1998.
- [5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, April 2022.
- [6] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *Proceedings of the International Conference on Machine Learning (ICML)*, Baltimore, Maryland, USA, July 2022.
- [7] T. Yoneda, J. Fang, P. Li, H. Zhang, T. Jiang, S. Lin, B. Picker, D. Yunis, H. Mei, and M. R. Walter, "Statler: State-maintaining language models for embodied reasoning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024.
- [8] Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with ChatGPT," *IEEE Access*, vol. 11, pp. 55 748–55 754, June 2023.
- [9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, December 2020.
- [10] M. Nakano, Y. Hasegawa, K. Funakoshi, J. Takeuchi, T. Torii, K. Nakadai, N. Kanda, K. Komatani, H. G. Okuno, and H. Tsujino, "A multi-expert model for dialogue and behavior control of conversational robots and agents," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 248–256, March 2011.
- [11] X. Xi, B. Xie, S. Zhu, T. Jin, J. Ren, and W. Song, "A general framework of task understanding for tour-guide robots in exhibition environments," in *Proceedings of the WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, Beijing, China, August 2022.
- [12] S. Hemachandra, T. Kollar, N. Roy, and S. Teller, "Following and interpreting narrated guided tours," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.
- [13] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The interactive museum tour-guide robot," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Madison, Wisconsin, USA, July 1998.
- [14] R. E. Fikes and N. J. Nilsson, "STRIPS: A new approach to the application of theorem proving to problem solving," *Artificial Intelligence*, vol. 2, no. 3, pp. 189–208, 1971.
- [15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [16] P. Khandelwal, S. Zhang, J. Sinapov, M. Leonetti, J. Thomason, F. Yang, I. Gori, M. Svetlik, P. Khante, V. Lifschitz, J. K. Aggarwal, R. Mooney, and P. Stone, "BWIBots: A platform for bridging the gap between ai and human-robot interaction research," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 5–7, February 2017.