

Few-Shot Inference of Human Perceptions of Robot Performance in Social Navigation Scenarios

Qiping Zhang

Yale University

New Haven, CT, USA

qiping.zhang@yale.edu

Nathan Tsoi

Yale University

New Haven, CT, USA

nathan.tsoi@yale.edu

Mofeed Nagib

Yale University

New Haven, CT, USA

mofeed.nagib@yale.edu

Hao-Tien Lewis Chiang

Google DeepMind

Mountain View, CA, USA

lewispro@google.com

Marynel Vázquez

Yale University

New Haven, CT, USA

marynel.vazquez@yale.edu

Abstract

Understanding how humans evaluate robot behavior during human-robot interactions is crucial for developing socially aware robots that behave according to human expectations. While the traditional approach to capturing these evaluations is to conduct a user study, recent work has proposed utilizing machine learning instead. However, existing data-driven methods require large amounts of labeled data, which limits their use in practice. To address this gap, we propose leveraging the few-shot learning capabilities of Large Language Models (LLMs) to improve how well a robot can predict a user's perception of its performance, and study this idea experimentally in social navigation tasks. To this end, we extend the SEAN TOGETHER dataset with additional real-world human-robot navigation episodes and participant feedback. Using this augmented dataset, we evaluate the ability of several LLMs to predict human perceptions of robot performance from a small number of in-context examples, based on observed spatio-temporal cues of the robot and surrounding human motion. Our results demonstrate that LLMs can match or exceed the performance of traditional supervised learning models while requiring an order of magnitude fewer labeled instances. We further show that prediction performance can improve with more in-context examples, confirming the scalability of our approach. Additionally, we investigate what kind of sensor-based information an LLM relies on to make these inferences by conducting an ablation study on the input features considered for performance prediction. Finally, we explore the novel application of personalized examples for in-context learning, i.e., drawn from the same user being evaluated, finding that they further enhance prediction accuracy. This work paves the path to improving robot behavior in a scalable manner through user-centered feedback.

CCS Concepts

- Human-centered computing → Social navigation; • Computing methodologies → Theory of mind; • Computer systems organization → Robotics.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

© Copyright held by the owner/author(s).

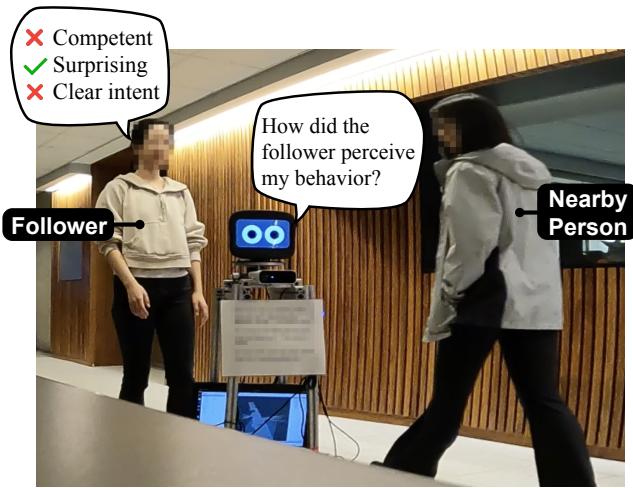


Figure 1: We investigate to what extent Large Language Models (LLMs) can infer human perceptions of a mobile robot in navigation scenarios where a person – the “follower” – was guided by the robot to an indoor location. The inferences are made based on a few examples only using In-Context Learning (ICL). For each example, the input consists of sensor-based observations from the robot and the output is a binary performance level (e.g., indicating competent behavior).

Keywords

human-robot interaction, large language models, few-shot learning

1 Introduction

Inferring how humans perceive a robot's performance is essential for designing robots that behave not only competently, but also in socially appropriate ways. These perceptions influence how people trust, collaborate with, and respond to robots in real-world settings [5, 12, 36, 40, 41, 51, 53]. Because people's perceptions of robot behavior are internal to the individual, they are typically measured through surveys that ask them to reflect on their experience. These evaluations tend to consider how the robot behaves in terms of different subjective factors that matter to humans, like whether the robot's actions are predictable or intentional [20, 21, 38]. However,

querying people during an interaction to assess their subjective perceptions of robot behavior can be disruptive and impractical [64, 65].

Prior work proposed to use supervised learning to infer human perceptions of robots from observable interaction data [64, 65]. First, survey data was collected via a user study, e.g., indicating how competent a person perceived the robot. Then, the data is used to train a learning model from scratch, such as a random forest or neural network. Once trained, the learned model can be used during interactions without having to query people via surveys again. This approach enabled more scalable robot behavior evaluation, although it required a significant data collection effort.

To make the learning approach more practical, we propose using Large Language Models (LLMs) to predict how a person perceives a robot’s behavior. Because LLMs encode world knowledge and have general reasoning capabilities [7, 15], we hypothesize that they can enable more efficient learning of human perceptions of robots than traditional supervised learning methods. Specifically, we investigate using In-Context Learning (ICL) to condition a pretrained LLM on a few labeled examples and prompt it to infer a user’s perception of a robot. This approach requires no retraining of the LLM, making it suitable for use in Human-Robot Interaction (HRI), where data tends to be limited and retraining of large models can easily result in overfitting.

Our work is focused on evaluating LLMs in social robot navigation scenarios, as in Fig. 1. Unfortunately, real-world social robot navigation data with human evaluations of robots is limited. Thus, we augmented an existing real-world dataset called SEAN TOGETHER [65] with additional robot-guided navigation episodes. This augmentation expands the prior dataset from 235 interaction episodes collected from 45 participants to 404 episodes from 69 participants. For each episode, the dataset provides ground truth human perceptions of a mobile robot considering three subjective factors: whether the robot is perceived as competent, whether its behavior is surprising, and whether the robot’s intentions are clear during navigation. We refer to the new dataset as the SEAN TOGETHER v2 dataset.¹

While prior work explored inferring human perceptions of robots with LLMs based on a high-level narrative description of an interaction [10, 63], we investigate making predictions using a robot’s sensor-based observations of the interaction. For example, we provide an LLM with observed motion trajectories, each represented as a list of coordinates. Using this data streamlines the application of LLMs.

Through a series of systematic experiments, our work helps us understand to what extent LLMs with ICL can infer internal human states from spatial robot data. Firstly, we analyze how LLM predictions compare to traditional supervised models in accuracy and sample efficiency. Then, we conduct an ablation study on the input features considered by an LLM, providing insights about what kind of sensor-based information it uses to infer human perceptions of a mobile robot. Finally, we investigate how tailoring demonstration examples in ICL to an individual affects the LLM’s ability to infer their perceptions of robot behavior. This effort is motivated by evidence that adapting learning models in HRI to individual users can result in better prediction performance (e.g., [43]). To

the best of our knowledge, our work is the first to explore creating personalized predictions of perceived robot performance, bringing us closer to a future where robot behavior can be evaluated at scale from a more individualized perspective than possible in prior work.

2 Related Work

Intuitive Psychology. We draw inspiration from emerging studies on LLM’s intuitive psychology capabilities, i.e., their ability to reason about human beliefs, goals, and social behaviors. Recent evaluations of LLMs consider varied Theory-of-Mind tasks from psychology [27, 28, 42, 55, 57], showing promise but also revealing brittleness under small task variations. Evaluation suites such as CogBench [11] suggest that Chain-of-Thought (CoT) prompting [61] can enhance LLM reasoning. Moreover, reviews on user modeling with LLMs call for personalized interactive systems [52].

Inferring Human Perceptions of Robots. Understanding how humans perceive robot behavior is central to developing robots that are not only functional but also behave desirably. Prior work has demonstrated using subjective evaluations of robot behavior to assess robot policies [4, 14, 36, 41, 51] and improve robot behavior [5, 12, 40, 48, 53].

We focus on predicting human perceptions that are critical in social robot navigation [20, 21]: robot competence, surprisingness, and clear intent. Competence reflects the robot’s ability to perform its intended task effectively [2, 8, 37, 54]. Surprisingness captures how much a robot’s behavior deviates from user expectations [3, 6, 20]. Clear intent refers to how easily a human can infer the robot’s goal and direction of motion [17, 18, 45]. These dimensions have been shown to shape people’s ability to coordinate with robots and their overall experience. Other perceptions such as discomfort [8, 26] and safety [1, 44] are also relevant, but are left as future work.

Prior research showed that it is possible to use supervised learning to predict human perceptions of robots [4, 64, 65]. In particular, we build directly on the work by Zhang et al. [65], who trained data-driven models (like a random forest model) to predict how people perceive a mobile robot during navigation. Different to prior work, though, we investigate few-shot learning, e.g., we consider learning from 4 examples versus 200+ examples as in [65]. To achieve sample-efficiency, we propose to use LLMs for the inference task.

Other recent work in HRI explores using LLMs for zero-shot inference, e.g., to predict human trust towards a robot [63], identify socially-appropriate robot navigation paths [46], identify robot errors [31], predict whether robot actions are explicable or legible [56], and whether a robot acted fairly [10]. While zero-shot prompting is practical, our results suggest that providing a few examples to LLMs can improve how well they infer human perceptions of robots.

Few-Shot Learning with LLMs. In-context learning (ICL) consists of conditioning LLMs on demonstration examples at inference time, without modifying the model’s parameters [7, 15, 32, 39]. This makes ICL appealing for robotics, where adaptation to new situations and users is often required, and where full model fine-tuning – such as with LoRA [25] – can be impractical due to computational cost and latency. Thus, ICL has gained popularity for adapting robot behavior (e.g., [13, 58, 62]). To our knowledge, our work is the first to use ICL to infer human perceptions of robots.

¹Link to data and code omitted for blind review.

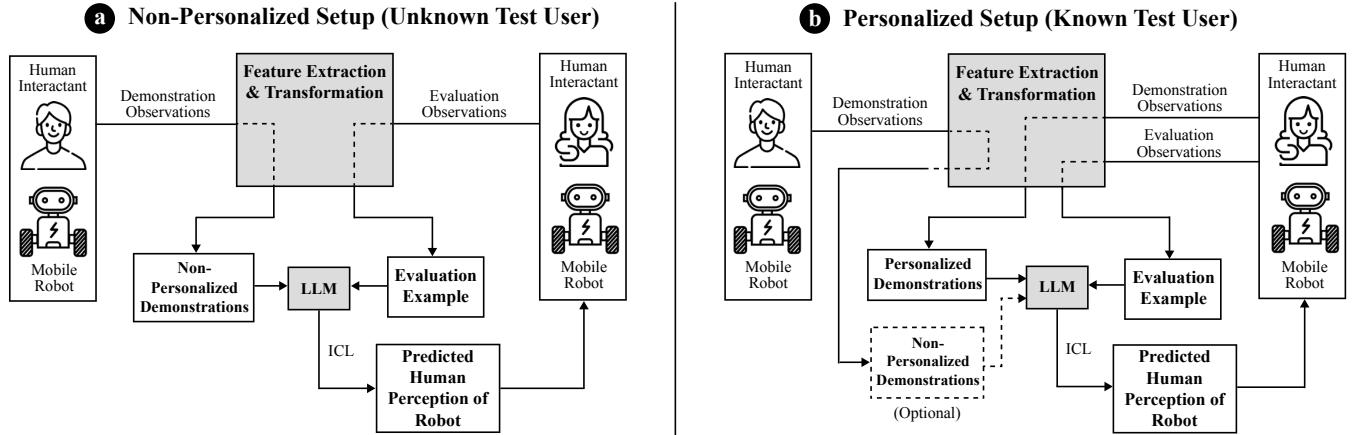


Figure 2: ICL overview: An LLM predicts a person’s perception of a robot on an *evaluation example* given a set of *demonstrations* in the prompt. In (a), demonstrations are gathered from interactions with users who are different from the person who generated the evaluation example. In (b), the demonstrations include examples from the same user who provided the evaluation example.

LLM-as-a-judge. Our work can be seen as an instance of the “LLM as a judge” paradigm, whereby an LLM is used to evaluate and assess the quality, relevance, or accuracy of outputs generated by other AI models. Prior work in machine learning has investigated whether LLMs with zero-shot or few-shot prompting can define rewards for Reinforcement Learning [29] and model user judgments about the behavior of an LLM [16, 30, 34]. For example, Kwon et al. [29] proposed prompting an LLM to assign reward values to an agent based on state-action trajectories from interaction scenarios. Their results demonstrate the feasibility of using LLMs to model user-aligned evaluations in games like the Ultimatum Game. Further, Dong et al. [16] proposed to personalize LLM predictions by providing a description of a persona, and Lau et al. [30] proposed to use the ICL capabilities of transformers to dynamically adapt LLM behavior to individual preferences in simulated human populations. Inspired by this work, we study the impact of ICL in HRI and evaluate personalized prompting in physical navigation tasks.

3 Method

We propose using LLMs with In-Context Learning to predict human perceptions of robots. The remaining of this section describes the ICL approach applied to a navigation scenario to facilitate the explanation; however, the same ICL formulation could be applied to other HRI interaction scenarios in the future.

Let $\mathcal{D} = \{(p^i, \mathbf{o}^i, y^i)\}_{i=1}^N$ be a dataset of human-robot interaction episodes, each having a finite time horizon T , that were collected when the robot interacted with a given person p^i . A sample (p^i, \mathbf{o}^i, y^i) has three values: the person index $p^i \in [1, P]$, a set of robot observations \mathbf{o}^i , and a perception label y^i provided by the person p^i in relation to the robot’s behavior. The label was collected at the end of the episode, when the person p^i completed a survey to provide their momentary perceptions of the robot. Following [65], we consider these perceptions as binary labels, e.g., a y^i indicates whether the robot behaved competently ($y^i = 1$) or not ($y^i = 0$) according to the person p^i . The observations \mathbf{o}^i are gathered by the robot during the interaction episode i and, for example, include

motion trajectories for the robot, the person p^i , and other nearby people over the time horizon T . Because our evaluation considers interactions where a robot guides the person p^i in an indoor environment, we refer to this person as the robot’s “follower” (Fig. 1).

3.1 ICL Setups

Fig. 2 illustrates the ICL setups that we consider in this work, which differ in terms of how prediction performance is measured. To explain the difference, consider an LLM \mathcal{M} and a previously-unseen *evaluation example* $(p^{eval}, \mathbf{o}^{eval}, y^{eval})$ from a test set generated from \mathcal{D} . The main goal of the LLM is to correctly predict the label y^{eval} based on a string representation of the robot observations \mathbf{o}^{eval} , which we refer to as the query $q = s(\mathbf{o}^{eval})$, with $s(\cdot)$ returning the string representation. To make a prediction, the LLM is additionally provided with the following information:

$$C = \{I, D_{\neq}, D_{\approx}\} \quad (1)$$

where I is the task instruction, and $D_{\neq} \in \mathcal{D}$ and $D_{\approx} \in \mathcal{D}$ are two different sets of demonstrations:

- *Non-personalized demonstrations.* The set D_{\neq} is gathered from interactions with users other than p^{eval} : $D_{\neq} = \{s(p^i, \mathbf{o}^i, y^i)\}_{i=1}^L$ where $p^i \neq p^{eval}$ and the function $s(\cdot)$ transforms the data to strings so the LLM can ingest it.

- *Personalized demonstrations.* The set D_{\approx} is gathered from interactions with the person p^{eval} : $D_{\approx} = \{s(p^i, \mathbf{o}^i, y^i)\}_{i=1}^M$ where $p^i = p^{eval}$ but $\mathbf{o}^i \neq \mathbf{o}^{eval}$. As before, $s(\cdot)$ transforms data to strings.

Then, C has $K = |D_{\neq}| + |D_{\approx}|$ demonstration examples in total. Finally, the LLM \mathcal{M} makes a prediction \hat{y} for the target y^{eval} as:

$$\hat{r} = \arg \max_{r \in \mathcal{R}} f_{\mathcal{M}}(r, C, q) \quad (2)$$

$$\hat{y} = \text{parse}(\hat{r}) \quad (3)$$

with $\hat{r} \in \mathcal{R}$ being the LLM’s string response, generated in an autoregressive manner with the model’s scoring function $f_{\mathcal{M}}$ [15]. The prediction \hat{y} is extracted from the response \hat{r} via a parser. In contrast

to supervised learning (as in [65]), the ICL formulation does not involve any model fine-tuning and relies solely on prompting.

In the **non-personalized ICL setup**, there are no demonstration examples that were collected from interacting with the person p^{eval} . This corresponds to making inferences when the set of personalized demonstration examples is *empty*, $C = \{I, D_{\neq}, \{\}\}$. Thus, the non-personalized setup can be seen as making predictions for an *unknown* test user, as in Fig. 2(a). This setup induces a data split similar to the supervised learning setups from prior work [64, 65].

Conversely, the **personalized ICL setup** corresponds to making predictions when the set of personalized demonstration examples is *not empty*. This can be seen as making predictions for a *known* test user, as illustrated in Fig. 2(b). We study two cases for the latter setup: in one case, only personalized examples are provided, so $C = \{I, \{\}, D_{\approx}\}$; in another case, both non-personalized and personalized examples are provided, so $C = \{I, D_{\neq}, D_{\approx}\}$. Fig. 3 illustrates the prompt structure for the non-personalized and personalized setups.

3.2 Observation Space

Based on findings from prior work [65], we utilize spatial behavior features for predicting perceived robot performance in social navigation scenarios. These features can be computed by mobile robots using off-the-shelf approaches for people tracking (e.g., with Kinect sensors [47]) and for robot localization [22].

For a given example (p, o, y) , the observations o provide a temporally grounded and robot-centric view of the navigation scene, encoding how the robot, the participant, and others in the environment move over time. The observations o span an 8-second time horizon and are represented in a coordinate frame centered on the robot's pose at the initial timestep ($t = 0$). Temporally-varying data is sampled at 1 Hz. The observations are:

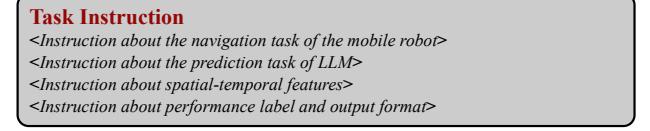
- *Goal Pose*: The 2D position of the robot's navigation goal relative to the robot at $t = 0$.
- *Robot Trajectory*: The robot's 2D position and orientation at each timestep of the time horizon.
- *Follower Trajectory*: The 2D position and orientation of the person following the robot at each timestep of the time horizon.
- *Nearby Pedestrians*: The 2D positions and orientations of other pedestrians at each timestep of the time horizon. We consider only observed people within a 7.2-meter radius of the robot, which corresponds to the robot's public space per Hall's proxemic zones [23].

All 2D positions are encoded as (x, y) locations, and the orientations θ are encoded as $(\cos \theta, \sin \theta)$. The cos-sin encoding is standard practice to ensure continuity for learning algorithms [50, 59, 64]. Although it is not as critical for LLMs, it helps supervised learning models, which we compare against in our evaluation.

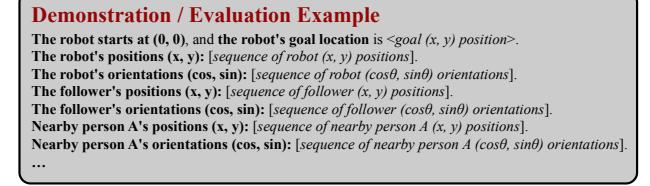
Fig. 3(b) illustrates how the observation features are included in the LLM's prompt. When a particular person is not detected in a timestep, their position and orientation are indicated as "unknown".

4 Experimental Setup

We evaluate ICL for predicting human perceptions of robot performance using a real-world HRI dataset, which is an augmentation of an existing dataset. Zhang et al. [65] contributed the SEAN TOGETHER dataset, which provides short episodes of human-robot interaction during a social navigation task in semi-public university



(a) Full prompt structure for ICL



(b) Prompt for a demonstration / evaluation example of the navigation task

Figure 3: Prompt structure (a), including the structure for an example (b). The LLM is asked to predict robot competence.

environments. Each episode contained observations of the interaction captured by the robot and corresponding human perceptions. We expanded the dataset with approval of our local Review Board using the same protocol and mobile robot. This increased the number of participants from 45 people to 69 people, totaling 404 labeled interaction episodes. We named the augmented dataset SEAN TOGETHER v2. See the supplementary video for example episodes.

Data Collection Protocol. The robot, which can be seen in Fig. 1, was built on a Pioneer 3-DX base. It was equipped with two Kinect sensors (one looking forward and one backwards), and had a screen-based face. It navigated autonomously through two public indoor spaces on a university campus. One space was a pedestrian tunnel; the other was a building entrance corridor.

The participants were not pre-recruited; rather, they engaged with the robot opportunistically. As pedestrians encountered the robot at the university, experimenters invited them to briefly follow the robot to a nearby goal marked on the ground. The robot periodically stopped during navigation and prompted the participants to evaluate its behavior using a mobile interface.

The robot's high-level behavior was implemented as in [65], where it either moved efficiently toward the destination (*Nav-Stack* behavior), spun in place to appear confused (*Spinning* behavior), and moved away from the goal (*Wrong-Way* behavior). The robot switched between the high-level behaviors to maintain a consistent

rate of sub-optimal behavior. The behaviors were designed to elicit both positive and negative views of the robot while also avoiding participant boredom or significant confusion.

Each behavior was executed for a fixed duration of 20-40 secs. Shortly before or after a behavior change, there was a pause in which participants rated the robot's recent performance. They answered 3 questions on a 5-point Likert format: *"How competent was the robot at navigating?"* (**Competence** performance dimension); *"How surprising was the robot's navigation behavior?"* (**Surprise**); and *"How clear were the robot's intentions during navigation?"* (**Intention**). As in [65], we inferred human perceptions based on an 8-second observation window preceding the participant's response.

Performance Labels. Zhang et al. [65] showed that predicting human perceptions of robots in a 5-point scale is very difficult even for humans. Thus, we created the performance labels for our experiments by binarizing the human ratings provided by the participants during data collection. Specifically, we excluded neutral-labeled examples, which were rare, and mapped the remaining responses by converting ratings of 4 or 5 to positive labels, and ratings of 1 or 2 to negative labels. In total, we had 363 episodes for the prediction of *Competence*, 351 for *Surprise*, and 375 for *Intention*.

Evaluation Procedure. Each perception dimension (*Competence*, *Surprise*, and *Intention*) is a separate classification task. For each dimension, we partitioned the participants into disjoint sets: 40% for testing, 40% for training, and 20% for validation. The splits were fixed across all experiments and conditions. To ensure that each participant in the test set contributed usable evaluation data, we only included participants who have at least one example with a positive label, one with a negative label, and at least four additional examples for personalization analysis. From each of the test participants, we randomly selected one positive and one negative example as the evaluation examples, resulting in a balanced evaluation set. Thus, we measured performance with classification *accuracy*.

All experiments are repeated 25 times because we use randomized demonstration sampling, whereby the demonstration examples are randomly chosen from the training data. For each of the 25 runs, all the models utilize the same demonstration and evaluation examples, and make predictions based on the same feature representation for the observations (as in Sec. 3.2) to ensure a fair evaluation.

5 Evaluation

We systematically investigated four research questions (RQs) using a limited number of demonstrations that ranged from $K = 4$ to $K = 64$.² We used linear mixed model analyses estimated with REstricted Maximum Likelihood (REML) [24, 49] to evaluate accuracy for each performance dimension. The analyses considered Run ID as a random effect because we repeated the experiments 25 times with varying demonstrations. The independent variables varied per RQ.

5.1 Non-Personalized ICL with Different LLMs

Our first research question was:

RQ1: Do LLMs with ICL result in more accurate, non-personalized

²Because LLMs can be biased by the distribution of target labels in the demonstration examples for ICL [60], we always set K to be a power of 2 so that we could balance the number of positive and negative examples whenever possible.

predictions of human perceptions of robot performance in few-shot learning scenarios than more traditional supervised learning?

In prior work [65], a Random Forest (RF) supervised learning model provided state of the art performance for predicting human perceptions of a guide robot. Thus, we compared the RF model in the the SEAN TOGETHER v2 dataset against several LLMs: Gemini 2.0 Flash; GPT 4.1 mini; and Llama 3.2 90B. The first two LLMs are closed-source models, while the latter one is open-source. Because part of our motivation for these models is making inferences *during* interactions, we limited the set of LLMs to "non-reasoning" models that can produce predictions relatively quickly on the cloud, e.g., hundreds of tokens per second, by simply predicting one token at a time in an auto-regressive fashion. For each LLM model, we also considered two prompting strategies: one with Chain-of-Thought (CoT) reasoning [61] that asked the model to "Do it step by step and explain your answer"; the other involved no CoT reasoning.

While ICL provides the demonstration examples to the LLMs via their prompt, the same demonstrations are used to train the RF from scratch. We thus hypothesized that with fewer examples, the LLMs would do better than the RF. To test this idea, we compared results in two scenarios: having few demonstrations with $K = |D_{\neq}| = 4$; and having a larger number with $K = |D_{\neq}| = 64$, which approached the limit of the context window for Llama given our prompt (Fig. 3).

Results. Table 1 shows prediction accuracy on 25 runs with $K = 4$ and $K = 64$ demonstration examples. The best result for the RF was on Surprise with $K = 64$, where it provided close performance to Gemini; otherwise, the RF model underperformed Gemini and GPT.

We analyzed the accuracy results with linear mixed models, one per performance dimension. Each analysis considered Run ID as a random effect, and Number of Demonstrations ($K = 4$ and $K = 64$) and Model (7 levels, each row of Table 1) as main effects. Also, the analysis considered the interaction between the main effects.

The analysis indicated that the Number of Demonstrations (K) had a significant effect on the accuracy for *Competence* ($F(1, 312) = 5.94, p = 0.0154$) and *Surprise* ($F(1, 312) = 53.38, p < 0.0001$). For *Competence*, a post-hoc Student's t-test showed that 64 demonstrations ($M = 0.65, SE = 0.007$) led to significantly higher accuracy than 4 demonstrations ($M = 0.64, SE = 0.006$) – although the average difference was close enough that it lacked functional meaning. A more pronounced significant difference was obtained for *Surprise*, where $K = 64$ led to an average accuracy of $M = 0.64 (SE = 0.01)$, and $K = 4$ led to $M = 0.60 (SE = 0.01)$. We attribute the limited effect of the K demonstrations on accuracy (considering several LLMs) to the challenge of processing long context windows [33]. For example, for Gemini No CoT with $K = 4$, the context window had about 3,000 tokens, while $K = 64$ led to about 36,000 tokens.

Model had a significant effect on the prediction accuracy ($p < 0.0001$ for all performance dimensions). Fig. 4 shows significant pairwise differences with Tukey HSD post-hoc tests. In general, the Gemini and GPT models resulted in significantly higher accuracy than Llama and RF. Although there were some significant differences in accuracy between the Gemini and GPT models, differences were small, showing the generalizability of our ICL approach.

Lastly, we found that the interaction between K and Model had a significant effect on accuracy, with $p < 0.0001$ for all performance dimensions. For $K = 4$, the Gemini and GPT models resulted in

Table 1: Results for RQ1. Average accuracy (\pm std. err.) of LLMs with ICL and Random Forest (RF) over 25 repetitions. CoT stands for Chain-of-Thought prompting. The **Best**, **Second**, and **Third** average results are highlighted.

Model	CoT	# Demo. Examples (K)	Competence	Surprise	Intention
Gemini 2.0 Flash	No	4	0.67 ± 0.01	0.65 ± 0.01	0.65 ± 0.02
Gemini 2.0 Flash	Yes	4	0.72 ± 0.01	0.64 ± 0.01	0.69 ± 0.01
GPT 4.1 mini	No	4	0.67 ± 0.01	0.64 ± 0.01	0.65 ± 0.01
GPT 4.1 mini	Yes	4	0.69 ± 0.01	0.64 ± 0.01	0.68 ± 0.01
Llama 3.2 90B	No	4	0.57 ± 0.01	0.55 ± 0.01	0.51 ± 0.01
Llama 3.2 90B	Yes	4	0.61 ± 0.01	0.52 ± 0.01	0.57 ± 0.01
RF	/	4	0.53 ± 0.02	0.57 ± 0.01	0.49 ± 0.02
Gemini 2.0 Flash	No	64	0.72 ± 0.01	0.70 ± 0.01	0.67 ± 0.01
Gemini 2.0 Flash	Yes	64	0.73 ± 0.01	0.67 ± 0.01	0.67 ± 0.01
GPT 4.1 mini	No	64	0.71 ± 0.01	0.67 ± 0.01	0.65 ± 0.01
GPT 4.1 mini	Yes	64	0.70 ± 0.01	0.66 ± 0.01	0.64 ± 0.01
Llama 3.2 90B	No	64	0.53 ± 0.01	0.59 ± 0.01	0.45 ± 0.01
Llama 3.2 90B	Yes	64	0.51 ± 0.01	0.48 ± 0.01	0.51 ± 0.01
RF	/	64	0.66 ± 0.01	0.69 ± 0.01	0.61 ± 0.02

significantly higher accuracy than RF and Llama in all dimensions. For $K = 64$, the post-hoc interaction tests showed that the Llama models resulted in significantly lower accuracy, but other results varied by dimension: for *Competence*, Gemini CoT and No CoT led to significantly higher accuracy than the other models, except for the GPT models; for *Surprise*, there was no significant difference for the Gemini, GPT, and RF models; and for *Intention*, Gemini No CoT had significantly higher accuracy than RF. In addition, for *Competence*, Gemini CoT with only $K = 4$ led to significantly higher accuracy than RF and Llama with $K = 64$, while for *Intention*, Gemini CoT and GPT CoT with only $K = 4$ led to significantly higher accuracy than RF and Llama with $K = 64$.

5.2 The Value of Spatial Observations for ICL

Our second research question was:

RQ2: Which spatial observations drove ICL performance with limited demonstrations ($K = 4$)?

We compared making non-personalized ICL predictions in a few-shot ICL scenario utilizing different types of observations: 1) the goal and robot trajectory only; 2) the goal, robot trajectory, and follower trajectory; and 3) the goal, robot trajectory, follower trajectory, and other pedestrian trajectories, i.e., all observations in Sec. 3.2. Thus, this RQ served as a feature ablation for ICL.

We limited our evaluation (and the following RQs) to Gemini 2.0 Flash with no CoT given our prior results, which showed strong performance for this model. Focusing on one model also helped reduce the cost of experiments and their carbon footprint [19].

Results. Fig. 5 shows prediction accuracy. For each performance dimension, we fit a linear mixed model on accuracy considering the Set of Observations provided to Gemini as main effect, and Run ID as random effect. The Set of Observations had a significant effect on *Competence* ($p < 0.0001$), *Surprise* ($p < 0.0001$), and *Intention* ($p = 0.0007$). Due to limited space, we summarize significant pairwise differences from Tukey HSD post-hoc tests in Fig. 5. The

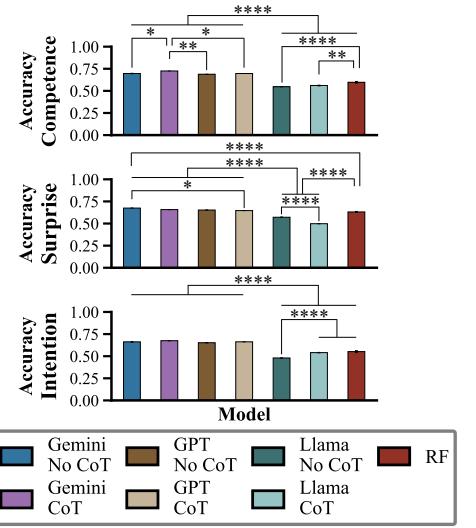


Figure 4: Model accuracy for RQ1. (****), (**) and (*) denote $p < 0.0001$, $p < 0.01$, and $p < 0.05$. Error bars are std. err. and are small.

results show that the LLM’s performance was not only due to using information about the robot trajectory and goal, but also to using pedestrian observations (including observations of the follower).

5.3 Increasing ICL Demonstrations

Our third research question was:

RQ3: How does ICL performance vary with an increasing number of demonstration examples (K) in the non-personalized setup?

We analyzed in more detail the impact of K on the ICL predictions for Gemini 2.0 Flash with no CoT, which provided good performance in RQ1. We considered $K \in \{4, 8, 16, 32, 64\}$, and two supervised learning baselines trained from scratch: RF (as in RQ1), and a neural network with a GRU architecture [9] (as in [64]). Also, we considered two other baseline models that required no training. First, we compared results with a weighted random sampling model (WR) that predicted a label by sampling from the distribution of targets in the demonstrations. This helped understand the complexity of the prediction problem. Further, we evaluated Gemini 2.0 Flash in a zero-shot prediction scenario, where the task instruction was the same as for ICL but no demonstrations were provided to the LLM (thus $K = 0$). This helped gauge how much the demonstrations contributed to the LLM’s performance given its world knowledge.

Results for Gemini only. The blue bars and line in Fig. 6 show average accuracy with an increasing K for Gemini. For each performance dimension, we analyzed accuracy using a linear mixed model with Run ID as a random effect, and Number of Demonstrations ($K \in \{0, 4, 8, 16, 32, 64\}$) as a main effects. The analysis showed a significant effect for K on all performance dimensions ($p < .0001$). For *Competence*, a Tukey HSD post-hoc test showed that the zero-shot model ($K = 0$) led to significantly lower accuracy than all ICL models ($K > 0$). Also, ICL with $K \in \{64, 16\}$ had significantly higher accuracy than ICL with $K = 4$. For *Surprise*, zero-shot Gemini also led to significantly lower accuracy than ICL. Further, ICL with $K = 64$ led to significantly higher accuracy than

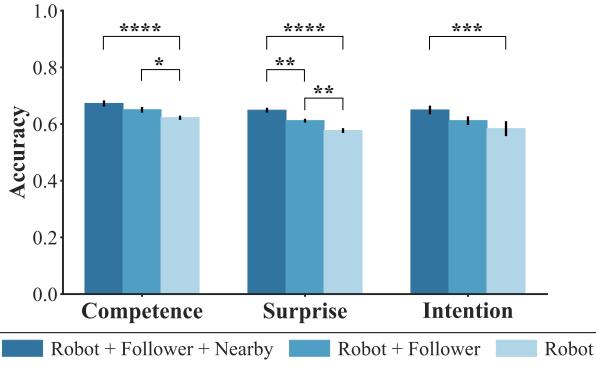


Figure 5: Results for RQ2. Average accuracy for Gemini 2.0 Flash No CoT with $K = 4$. The model always takes as input the goal location, but the other spatial observations are ablated. Error bars are std. err. The symbols (****), (***) and (**) denote $p < 0.0001$, $p < 0.001$, $p < 0.01$, and $p < 0.05$.

$K \in \{16, 8, 4\}$. For *Intention*, ICL with $K \in \{32, 16\}$ led to significantly higher accuracy than $K = 4$ and the zero-shot model. Taken together, these results suggest that the LLM benefited from having demonstrations.

Results for Gemini vs Other Models. Fig. 6 shows results for all the models. We analyzed prediction accuracy on each performance dimension using a linear mixed model, but this time excluded the zero-shot case for which the supervised learning models could not be fit. The linear mixed model considered Run ID as a random effect, Number of Demonstrations ($K \in \{4, 8, 16, 32, 64\}$) and Model (*Gemini*, *RF*, *GRU*, *WR*) as a main effects, and the interaction between the Number of Demonstrations and Model. Because of limited space, we focus on discussing the interaction effect, which is the most relevant for RQ3 and was significant for *Competence* ($p = 0.0275$), *Surprise* ($p < 0.0001$), and *Intention* ($p = 0.0036$).

The Tukey HSD post-hoc tests for the interaction effect confirmed the superiority of the Gemini model in most cases. For example, for *Competence*, Gemini with $K \in \{64, 32, 16, 8\}$ led to significantly higher accuracy than all other models, except for Gemini with $K = 4$ and RF with $K = 64$. In contrast to our prior results for RQ3 considering Gemini only, the post-hoc tests for the interaction effect between the Number of Demonstrations and Model resulted in no significant pairwise differences for Gemini across $K = \{4, 8, 16, 32, 64\}$. RF benefited more from an increasing K . The RF model with $K = 64$ demonstrations led to significantly higher accuracy than RF with $K = \{4, 8\}$ across all performance dimensions.

5.4 ICL with Personalized Demonstrations

Our last research question was:

RQ4: Do personalized examples improve in-context learning?

Because individual factors can influence human perceptions of robots, we examined whether we could improve ICL accuracy with prompts constructed with demonstration examples from the same user who provides the evaluation example (as in Fig. 2(b)). Following RQ2 and RQ3, we considered only Gemini 2.0 Flash with no CoT for this experiment. Also, we considered three values for the

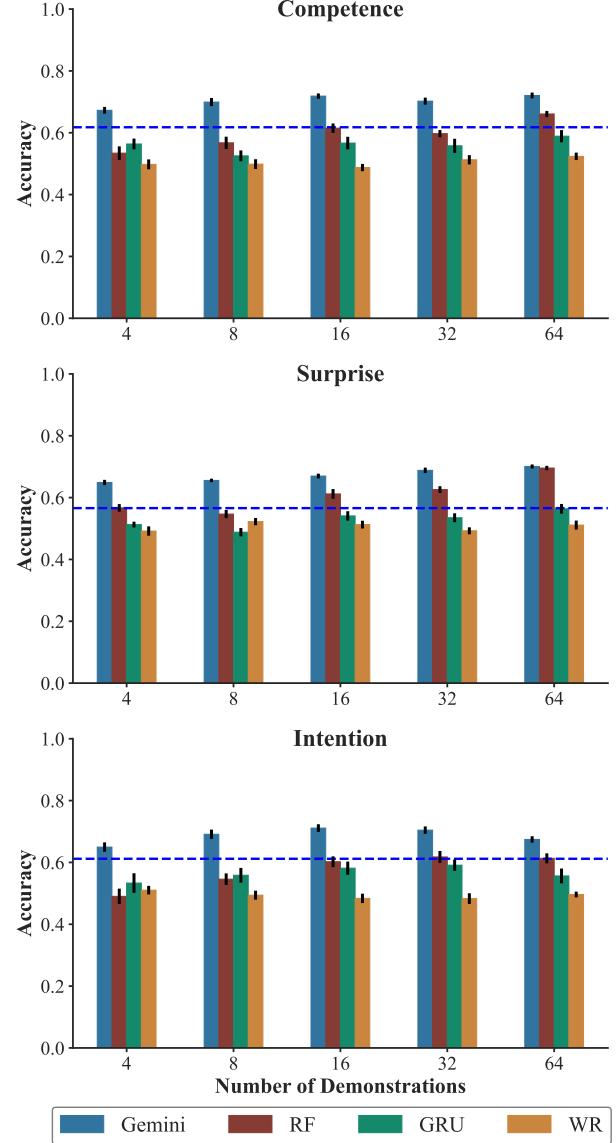


Figure 6: Results for RQ3. Accuracy of Gemini 2.0 Flash No CoT (*Gemini*), Random Forest (*RF*), Recurrent Network (*GRU*), and Weighted Random Sampling (*WR*) with varying number of demonstrations (K). The blue line indicates average accuracy for Gemini 2.0 Flash with a zero-shot prompt.

total number of ICL demonstrations K : 4, 8 and 68 total examples. Specifically, for each K value, we either had:

- 0 personalized demonstrations (so $|D_{\sim}| = 0$ and $K = |D_{\neq}|$); or
- 4 personalized demonstrations (so $|D_{\sim}| = 4$ and $K = |D_{\neq}| + 4$).

Results. The results are shown in Table 2. For each performance dimension, we analyzed accuracy using a linear mixed model with Run ID as a random effect, the Number of Personalized Demonstrations ($|D_{\sim}| \in \{0, 4\}$) and the Total Number of Demonstrations ($K \in \{4, 8, 68\}$) as the main effects, and their pairwise interaction.

Table 2: Results for RQ4. Mean accuracy (\pm std. err.) of Gemini 2.0 Flash with varying numbers of personalized and non-personalized demonstrations, over 25 repetitions. The **Best, **Second**, and **Third** average results are highlighted.**

Row	Num. of Non-Personalized Demonstrations ($ D_{\neq} $)	Num. of Personalized Demonstrations ($ D_{\approx} $)	Total Number of Demonstrations (K)	Competence	Surprise	Intention
1	0	4	4	0.71 ± 0.01	0.69 ± 0.01	0.72 ± 0.00
2	4	0	4	0.67 ± 0.01	0.65 ± 0.01	0.65 ± 0.02
3	4	4	8	0.76 ± 0.01	0.73 ± 0.01	0.76 ± 0.01
4	8	0	8	0.68 ± 0.01	0.65 ± 0.01	0.64 ± 0.02
5	64	4	68	0.79 ± 0.01	0.76 ± 0.00	0.75 ± 0.01
6	68	0	68	0.72 ± 0.01	0.70 ± 0.01	0.69 ± 0.01

The Number of Personalized Demonstrations ($|D_{\approx}|$) had a significant effect on accuracy on all performance dimensions ($p < 0.0001$). The post-hoc test showed that using 4 personalized demonstrations led to significantly higher accuracy than using zero ($|D_{\approx}| = 0$).

Additionally, the Total Number of Demonstrations (K) had a significant effect on the accuracy. For *Competence* and *Surprise* ($p < 0.0001$), Tukey HSD post-hoc tests indicated that $K = 68$ total examples led to significantly higher accuracy than $K = 8$, which also led to significantly higher accuracy than $K = 4$. For *Intention* ($p = 0.0052$), $K = 68$ led to significantly higher accuracy than $K = 4$.

Lastly, we found that the interaction between $|D_{\approx}|$ and K had a significant effect on accuracy for *Surprise* ($p = 0.0258$) and *Intention* ($p = 0.0364$), but not *Competence*. For *Surprise*, using $K = 68$ or $K = 8$ demonstrations, including $|D_{\approx}| = 4$ examples, led to significantly higher accuracy than the other options in Table 2. For *Intention*, using $K = 68$ and $|D_{\approx}| = 4$, or using $K = 8$ and $|D_{\approx}| = 4$, led to significantly higher accuracy than the other options except for using only $K = |D_{\approx}| = 4$ personalized demonstrations. We conclude that the personalized demonstrations helped ICL performance.

Qualitative Analysis. To better understand ICL performance, we manually inspected predictions by Gemini 2.0 Flash No CoT with $|D_{\neq}| = 64$ and $|D_{\approx}| = 4$, for a total of 68 demonstrations. As shown in row 5 of Table 2, this model had highest average accuracy on *Competence* and *Surprise*, and was second best for *Intention*.

For each perception dimension, we selected three sets of 10 examples from the test set: the 10 examples with the highest prediction accuracy, the 10 with the lowest, and the 10 with accuracy closest to 50% across the 25 runs of the model. Then, we visualized the navigation episodes and identified recurring patterns that correlated with the model’s success, failure, or prediction ambiguity. Unsurprisingly, the model achieved high accuracy on episodes with consistent robot behaviors, such as steady progress towards the goal or aimless rotation far from it. Low and middle-accuracy predictions mainly stemmed from:

1) *Semantic ambiguity* of the robot’s final state (6/20 examples for *Competence*; 4/20 examples for *Surprise*; and 8/20 examples for *Intention*). We did not provide the LLM a specific threshold for when the robot reached the goal, which made it difficult to gauge how close was close enough to complete the navigation task.

2) *Transitional ambiguity* from mid-episode behavioral shifts, such as corrective turns or reversals (5/20 examples for *Competence*, and 6/20 examples for *Surprise* and *Intention*). The robot showed both effective and ineffective behavior within an episode.

3) *Contextual and kinematic ambiguity* in the robot behaviors, like navigating away from the goal or rotating in place (7/20 examples for *Competence*, 6/20 for *Surprise*, and 7/20 for *Intention*). Model uncertainty seemed to stem from subtle variations in the context (e.g., proximity to the goal) or kinematics (e.g., a slow drift vs. a rapid retreat) that were under-sampled in the demonstrations.

6 Discussion

Summary of Key Findings. We proposed an In-Context Learning approach for LLMs to infer human perceptions of robot performance. The approach used observations of interactions to predict user evaluations of robot competence, surprisingness, and intent. Our experiments in navigation scenarios showed that our ICL approach not only matches or exceeds the performance of traditional supervised models with a fraction of the data but also outperforms zero-shot LLM predictions. Further, accuracy is enhanced by personalizing in-context examples to the test user.

ICL Limitations. Despite its promise, our work also highlighted limitations of ICL. For example, we found mixed results on whether more demonstrations increased prediction accuracy. We suspect the root cause is that with more examples, the context window for the LLM is longer, which can make LLMs struggle [33]. Also, we investigated choosing demonstrations by random sampling, but perhaps a more thoughtful approach could help LLMs better leverage more demonstrations [35, 66]. Importantly, ICL required prompt engineering effort. For example, in early experiments, we found that unintuitive values for the target label could reduce performance, such as using “surprise = 0” to indicate surprising behavior. In a qualitative analysis, we also found that semantic ambiguity in what it meant to complete the navigation task in our prompt could lead to erroneous predictions. More systematic experiments are needed to assess the robustness of the ICL approach to prompt variations.

Future Work. Limitations of our research also point to future research directions. First, our evaluation focused on a specific robot-following task. More work is needed to validate ICL across more diverse interactions. Second, our observation representation was limited to spatio-temporal features provided as strings to LLMs. Incorporating multimodal cues (e.g., using videos captured from the robot) could increase performance with multi-modal large models. Ultimately, the most compelling application of this work is robot behavior improvement. A robot could use its predictions of user perceptions as direct feedback to adjust its behavior policy, closing the loop from passive inference to active, socially-aware adaptation.

Acknowledgements

Thanks to Kenneth Shui for helping with data collection. This work was partially supported by the National Science Foundation (NSF) under Grant No. IIS-2143109, and Google under a Gemini credits grant. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the NSF or Google.

References

- [1] Nezih Akalin, Annica Kristoffersson, and Amy Loutfi. 2022. Do you feel safe with your robot? Factors influencing perceived safety in human-robot interaction based on subjective and objective measures. *International journal of human-computer studies* 158 (2022), 102744.
- [2] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. 2022. You Are In My Way: Non-verbal Social Cues for Legible Robot Navigation Behaviors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 657–662.
- [3] Chatchalita Asavani and Hiroyuki Umemuro. 2021. Personal space violation by a robot: An application of expectation violation theory in human-robot interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1181–1188.
- [4] Pilar Bachiller-Burgos, Ulysses Bernardet, Luis V Calderita, Pranup Chhetri, Anthony Francis, Noriaki Hirose, Noé Pérez, Dhruv Shah, Phani T Singamaneni, Xuesu Xiao, et al. 2025. Towards Data-Driven Metrics for Social Robot Navigation Benchmarking. *arXiv preprint arXiv:2509.01251* (2025).
- [5] Aniket Bera, Tanmay Randhavane, and Dinesh Manocha. 2019. Improving Socially-aware Multi-channel Human Emotion Prediction for Robot Navigation. In *CVPR Workshops*. 21–27.
- [6] Martim Brandao, Gerard Canal, Senka Krivić, Paul Luff, and Amanda Coles. 2021. How experts explain motion planner output: a preliminary user-study to inform the design of explainable planners. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 299–306.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [8] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.
- [9] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- [10] Houston Claire, Aly Moosa, and Marynel Vázquez. 2025. Inferring Human Fairness Judgments with Large Language Models in Human-Robot Interaction Scenarios. In *2025 IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE.
- [11] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225* (2024).
- [12] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. 2021. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*.
- [13] Norman Di Palo and Edward Johns. [n. d.]. Keypoint Action Tokens Enable In-Context Imitation Learning in Robotics. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.
- [14] Antonio Di Tecco, Antonio Frisoli, and Claudio Loconsole. 2025. Machine Learning Prediction on User Satisfaction in Human-Robot Interaction (HRI) Tasks. *IEEE Access* (2025).
- [15] Qinxin Dong, Lei Li, Darnai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).
- [16] Yijiang Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a Personalized Judge?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10126–10141.
- [17] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 51–58.
- [18] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.
- [19] Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. [n. d.]. LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models. In *The Twelfth International Conference on Learning Representations*.
- [20] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. 2023. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740* (2023).
- [21] Yuxiang Gao and Chien-Ming Huang. 2022. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI* 8 (2022), 721317.
- [22] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. 2007. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE transactions on Robotics* 23, 1 (2007), 34–46.
- [23] Edmund T Hall and Edward T Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.
- [24] David A. Harville. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *J. Amer. Statist. Assoc.* 72, 358 (1977), 320–338. <http://www.jstor.org/stable/2286796>
- [25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
- [26] Hiroyuki Kidokoro, Takayuki Kanda, Dražen Bršić, and Masahiro Shiomi. 2013. Will I bother her? A robot anticipating its influence on pedestrian walking comfort. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 259–266.
- [27] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* 4 (2023), 169.
- [28] Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* 121, 45 (2024), e2405460121.
- [29] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001* (2023).
- [30] Allison Lau, Younwoo Choi, Vahid Balazadeh, Keertana Chidambaram, Vasilis Syrigkanis, and Rahul G Krishnan. 2024. Personalized adaptation via in-context preference learning. *arXiv preprint arXiv:2410.14001* (2024).
- [31] Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. 2025. The Human Robot Social Interaction (HSRI) Dataset: Benchmarking Foundational Models' Social Reasoning. *arXiv preprint arXiv:2504.13898* (2025).
- [32] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).
- [33] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060* (2024).
- [34] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yéjin Choi. [n. d.]. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *The Twelfth International Conference on Learning Representations*.
- [35] Jichang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 100–114.
- [36] Shih-Yun Lo, Katsu Yamane, and Ken-ichiyo Sugiyama. 2019. Perception of pedestrian avoidance strategies of a self-balancing mobile robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1243–1250.
- [37] Christoforos Mavrogiannis, Patricia Alves-Oliveira, Wil Thomason, and Ross A Knepper. 2022. Social momentum: Design and evaluation of a framework for socially competent robot navigation. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 2 (2022), 1–37.
- [38] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. 2023. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–39.
- [39] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721* (2023).
- [40] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Adapting robot behavior for human-robot interaction. *IEEE Transactions on Robotics* 24, 4 (2008), 911–916.
- [41] Sören Pirk, Edward Lee, Xuesu Xiao, Leila Takayama, Anthony Francis, and Alexander Toshev. 2022. A protocol for validating social navigation policies. *arXiv preprint arXiv:2204.05443* (2022).
- [42] Sushrita Rakshit, James Hale, Kushal Chawla, Jeanne M. Brett, and Jonathan Gratch. 2025. Emotionally-Aware Agents for Dispute Resolution. In *Proceedings of the 13th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Canberra, Australia.
- [43] Silvia Rossi, François Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99 (2017), 3–12.

[44] Matteo Rubagotti, Inara Tusseyeva, Sara Baltabayeva, Danna Summers, and Anara Sandygulova. 2022. Perceived safety in physical human–robot interaction—A survey. *Robotics and Autonomous Systems* 151 (2022), 104047.

[45] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. 2018. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine* 37, 1 (2018), 22–29.

[46] Zhonghao Shi, Enyu Zhao, Nathaniel Dennler, Jingzhen Wang, Xinyang Xu, Kaley Shrestha, Mengxue Fu, Daniel Seita, and Maja Matarić. 2025. HRIBench: Benchmarking Vision-Language Models for Real-Time Human Perception in Human-Robot Interaction. *arXiv preprint arXiv:2506.20566* (2025).

[47] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.

[48] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. 2024. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters* (2024).

[49] Walter W Stroup. 2012. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

[50] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. 2020. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[51] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. 2019. From one to another: how robot-robot interaction affects users’ perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 114–122.

[52] Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518* (2023).

[53] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[54] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, JD Zhao, and Marynel Vázquez. 2021. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7528–7535.

[55] Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

[56] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of mind abilities of large language models in human-robot interaction: An illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 36–45.

[57] Lennart Wachowiak, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. Are Large Language Models Aligned with People’s Social Intuitions for Human-Robot Interactions?. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2520–2527.

[58] Peng-Yuan Wang, Jing-Cheng Pang, Chen-Yang Wang, Xuhui Liu, Tian-Shuo Liu, Si-Hang Yang, Hong Qian, and Yang Yu. 2025. InCLET: Large Language Model In-context Learning can Improve Embodied Instruction-following. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2134–2142.

[59] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. 2017. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2043–2050.

[60] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems* 36 (2023), 15614–15638.

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[62] Yida Yin, Zekai Wang, Yuvan Sharma, Dantong Niu, Trevor Darrell, and Roei Herzig. 2025. In-context learning enables robot action prediction in llms. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8972–8979.

[63] Bowen Zhang and Harold Soh. 2023. Large language models as zero-shot human models for human-robot interaction. In *2023 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 7961–7968.

[64] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 398–407.

[65] Qiping Zhang, Nathan Tsoi, Mofeed Nagib, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, and Marynel Vázquez. 2025. Predicting Human Perceptions of Robot Performance during Navigation Tasks. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–27.

[66] Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9134–9148.