2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Self-Confidence from Semantic Action Embeddings for Improved Trust in Human-Robot Interaction

Cedric Goubard[1] and Yiannis Demiris[1]

***Abstract*— In Human-Robot Interaction (HRI) scenarios, human factors like trust can greatly impact task performance and interaction quality. Recent research has confirmed that perceived robot proficiency is a major antecedent of trust. By making robots aware of their capabilities, we can allow them to choose when to perform low-confidence actions, thus actively controlling the risk of trust reduction. In this paper, we propose Self-Confidence through Observed Novel Experiences (SCONE), a policy to learn self-confidence from experience using semantic action embeddings. Using an assistive cooking setting, we show that the semantic aspect allows SCONE to learn self-confidence faster than existing approaches, while also achieving promising performance in simple instructions following. Finally, we share results from a pilot study with 31 participants, showing that such a self-confidence-aware policy increases capability-based human trust.**

I. INTRODUCTION

Trust is a key element of any Human-Robot Interaction (HRI), as robots can help achieve a better interaction quality by gauging a human's trust and adapting their actions accordingly [1]. An important aspect of effective collaboration is understanding when to intervene, as failures not only inherently bring risks, but also affect the human's future trust [2]. This requires robots to be aware of their own capabilities, which is known as proficiency self-assessment [3]. This is a difficult task: to achieve it, robots must be able to learn proficiency models quickly, accurately and generalise them to various contexts. To that purpose, special care must be given to the way actions are represented and understood by the robot. In adjacent fields, the use of language-anchored action embeddings has recently yielded promising results thanks to their ability to capture functional similarities between actions [4], [5], [6]. In this paper, we consider how this idea can extend to proficiency self-assessment and the resulting impact on human trust levels during an interaction.

We choose assistive cooking as our application scenario, with the robot in charge of providing the right ingredients at the right time to the human. There are several reasons why this choice fits our problem particularly well. First, it is a realistic example of a complex HRI scenario. Second, it encompasses a wide range of possible robot actions, which forces our approach to deal with the problems of exploding state and action spaces. Additionally, it includes several types of risks, a key element when investigating trust in HRI [7]. Examples include breaking objects, wasting food or time, spilling, or staining. Finally, assistive cooking allows us to evaluate embeddings for both a performance-oriented task (instructions following) and self-confidence modelling.

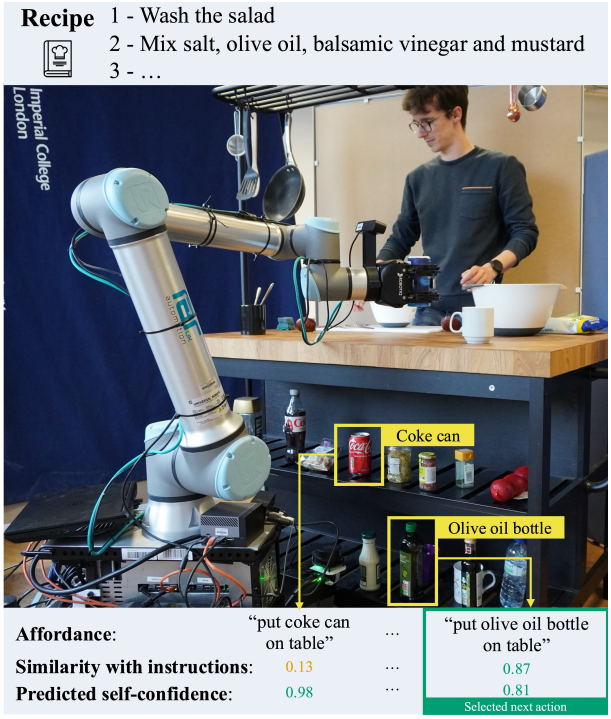


Fig. 1: By using semantic embeddings to represent available actions, the robot can predict a similarity with a set of instructions and a success chance based on past experience. The figure shows this approach applied in our assistive cooking setup; predictions are in the bottom, instructions in the top part. The robot can thus ask for help before failures happen, resulting in a positive effect on capability-based human trust.

In this paper, we present Self-Confidence through Observed Novel Experiences (SCONE), a framework to learn a self-confidence-informed policy that also allows simple instructions following. Our contributions are as follows:

1) We present the details of SCONE and how we use a shared embedding space to learn self-confidence from experience and follow simple instructions.
2) We evaluate the performance of several embedding spaces and predictive models on both the success prediction and instructions following tasks.
3) We investigate SCONE's effect on reported and behavioural trust through a pilot study involving a real-life assistive cooking scenario.

[1]Authors are with the Personal Robotics Laboratory, Dept. of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ. This research was supported in part by UKRI grant EP/V026682/1, and by a Royal Academy of Engineering Chair in Emerging Technologies to Yiannis Demiris. {c.goubard21, y.demiris}@imperial.ac.uk

II. RELATED WORK

A. Robot self-confidence

Human overreliance on automated systems can lead to lower trust when these systems fail [8], [9]. To address this concern, researchers have explored methods to imbue robots with self-awareness and active engagement in the human trust calibration process. A prevalent approach involves utilising expected reward mechanisms [5], [10]. However, this approach is focused on Reinforcement Learning (RL)-based agents and proves difficult to implement within HRI scenarios due to the complexities of human behaviour modelling. As an alternative, comprehensive frameworks have been developed and exhibit promising outcomes [11], [12], [13], [14]. These models consider a wide range of factors impacting task proficiency, albeit with specific prerequisites that restrict their adaptability. These prerequisites include elements like reward structuring, exhaustive PDDL definitions, hierarchical databases to estimate task similarity, or stipulating cognitive assumptions of the agent. Our proposed approach offers a simpler method which can be easily applied to a wide range of tasks thanks to the versatility of text descriptions. We argue that this is a valuable aspect, as it can be more easily included in various types of HRI user studies. Moreover, despite the extensive research on robot self-confidence, we believe this work to be the first one to study the impact of a learnt, task-agnostic self-confidence on both reported and behavioural trust.

B. Instructions following

Strategies designed to allow robots to follow instructions can be broadly categorised into two distinct groups, depending on whether they use implicit or explicit action representations. Approaches using implicit action representations involve training models to directly generate robot commands from sensor data [15], [16]. However, these approaches require extensive simulation-based training and do not consider potential human actions in the process. As an alternative, explicit representations deconstruct long-term instructions into more progressive, short-term steps, including elements such as skill descriptions, latent actions, PDDL states, or vision/proprioception states [5], [17], [18], [19]. Furthermore, interesting progress has been made over the recent years by using Large Language Models (LLMs), partly thanks to their growing ability to emulate human reasoning and concepts. Such progress has affected both implicit [4], [5] and explicit approaches [20], [21], [22], [23]. While LLM-based embeddings have been extensively explored, we believe the work presented here to be the first to investigate their use for robot self-confidence estimation and its effect on trust in a realistic HRI.

C. Trust in Human-Robot Interaction

Trust is a subjective, context-dependant, and multidimensional phenomenon [24], [25]. In attempts to capture this complexity, several definitions have been proposed over the years [26], [27]. In this paper, we adopt a commonly accepted one in HRI: 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability' [28]. Additionally, recent research differentiates *capability-based trust*, *i. e.* related to the belief over what an agent is physically capable of doing, and *moral-based trust*, *i. e.* whether the agent's motive and morals are aligned with the human's [29], [30]. When attempting to measure trust, an important distinction must be noted between *reported trust*, mostly measured using questionnaires, and *behavioural trust*, observed for instance through delegation and intervention. Behavioural trust is the one directly affecting the HRI; when it is not easily available, reported trust can offer a useful but noisy proxy, as it does not always translate to behavioural trust [31], [32]. However, both measures have proven useful, and human trust estimation and calibration have been subject to extensive research over the past few years [30], [33], [34], [35], [36], [37], [38]. While this research provides valuable insight, we believe our work to be the first to study the impact of explicit, task-agnostic robot self-confidence on both reported and behavioural trust in a physical HRI setting.

III. THEORETICAL FRAMEWORK

A. Terminology

A robot's overall probability of successfully grasping a bottle is not equal to the probability of grasping a specific bottle on a cluttered desk. To emphasise this distinction between abstract and grounded actions, we isolate three distinct concepts. An agent's **capability** $\lambda \in \Lambda$ is a potential sequence of commands assigned to specific purposes, such as 'grasping' or 'navigating'. An **affordance** $\alpha \in A$ arises when a capability can be applied to an element of the environment $e \in \mathcal{E}$, such as 'the bottle on the desk' or 'the kitchen'. This is in line with the original definition of 'action possibilities that are offered by the environment' [39], as well as other work from the literature [40]. Finally, when an affordance is executed, this leads to an **outcome** ω_α obtained using an outcome function Ω. We will assume binary outcomes as they fit our application, but the framework could easily be generalised to continuous space. We also consider that an affordance exists as long as the robot could intentionally achieve it successfully at least once. Equation (1) summarises the mathematical notations.

$$\begin{aligned} \alpha = (\lambda, e) \in A = \Lambda \times \mathcal{E}, \\ \omega_\alpha = \Omega(\alpha) \in \{0, 1\}. \end{aligned} \tag{1}$$

$$\forall \lambda \in \Lambda,\ \kappa(\lambda) = \mathop{\mathbb{E}}_{e \in \mathcal{E}}[\Omega(\alpha)] = \mathop{\mathbb{E}}_{e \in \mathcal{E}}[\Omega(\lambda, e)] \in [0, 1]. \tag{2}$$

These clarifications now allow us to refine our definition of self-confidence. We define an agent's **proficiency** at a capability $\kappa(\lambda)$ as the expected success rate of all associated affordances over the distribution of contexts, as shown in Equation (2). The **self-confidence** is then simply the predicted proficiency $\hat{\kappa}$ based on past experiences. Our framework examines several methods to obtain this estimate in Section IV-A. Considering self-confidence as 'trust in oneself', we can relate our notations to existing work, such as [34], but with the Markovian assumption released, or [41], but using a latent capability space.

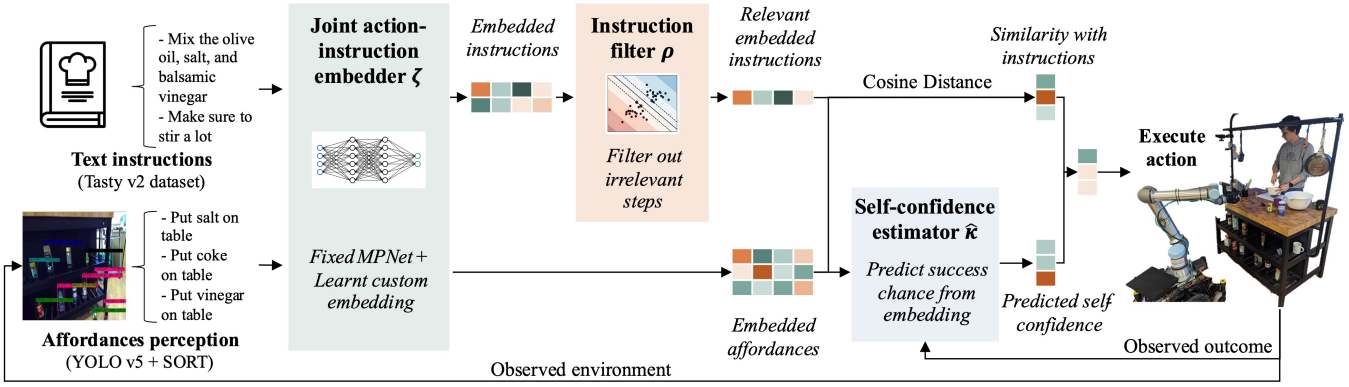


Fig. 2: **Our architecture applied to an assistive cooking scenario**. Affordances with textual descriptions are extracted from the environment using object tracking. They are embedded using ζ in a common space with the instructions, which are filtered by ρ to remove irrelevant instructions for the robot. Actions and instructions embeddings are then compared using the cosine distance. The same action embeddings are used by $\hat{\kappa}$ to predict their success chance, and the most relevant and safe action is picked. Its outcome is observed and used to update the self-confidence.

B. Framework Overview

First, we define the text space $\mathcal{T}$ and assume that a description function $\delta : A \to \mathcal{T}$ is available to get text descriptions of affordances; examples on how to get such a function are available in the literature [40], [42] . In our application scenario, δ is simply 'Put X on table' where X is the class name from the object detection module, and the instructions are the recipe steps. We propose SCONE, a policy framework relying on 3 main components, with implementation details shared in Section IV. First, an **action-instruction joint embedder** $\zeta : \mathcal{T} \to \hat{V}$ maps the textual description of the instructions and available affordances to a joint vector space $\hat{V}$. Having instructions and actions in a shared semantic space enables us to measure their similarity, and serves as a base for the other two modules. Then, an **instruction relevance filter** $\rho : \hat{V} \to \{0, 1\}$ uses the embeddings from ζ to (1) identify instructions that have been completed and (2) filter out irrelevant instructions for the robot, which reduces the complexity of the planning. The output is binary, specifying whether or not to keep the instruction. Finally, a **self-confidence estimator** $\hat{\kappa} : \hat{V} \to [0, 1]$ maps an action embedding to a success probability. Since $\hat{V}$ captures similarities between related actions, $\hat{\kappa}$ converges to the real proficiency much faster than other approaches such as [6] where the outcome of each action is only used to update that specific action, as shown in Section V-A. An implementation example in an assistive cooking scenario is available in Figure 2. To learn from experience, the model used as a self-confidence estimator (see Section IV-A) is updated after the outcome of each action is observed.

IV. EXPERIMENTS

Our application scenario is an assistive cooking setup where the robot delivers ingredients while the human performs the recipe steps. As mentioned in Section I, we chose this scenario because it is a complex HRI scenario with different risks involved and large action and state spaces. We conducted three sets of experiments, each designed to evaluate a specific dimension of our framework. Section IV-A focuses on the self-confidence learning process, evaluating how fast and accurate it is. Section IV-B considers how the same embeddings perform on our core task, instructions following. Finally, Section IV-C presents our pilot user study to measure the impact of self-confidence-awareness on human trust in a real-life assistive cooking scenario.

A. Experiment 1: Learning self-confidence

As described in Section III-B, the overall ability to predict an affordance's outcome depends both on the joint embedding module ζ and on the self-confidence estimator $\hat{\kappa}$. For the embedding step, we selected two semantic embeddings from the current state of the art, MPNet [43] and BGE [44], both accessed using the `sentence_transformers` package [45]. We also trained a custom version of MPNet with extra layers on a custom dataset (see Section IV-B), which we included in the experiment. As far as the authors know, there is no commonly accepted baseline for robot action embedding in a success prediction task, as most related work uses either a task-specific model or different representations (see Section II-A). We thus decided to adopt two baselines from adjacent problems: one-hot encodings and GloVe [46]. Both have been used to embed robot task descriptions in state-of-the-art work [6], [34], [40].

Embeddings alone are not enough: we also needed to compare estimators $\hat{\kappa}$ to predict the self-confidence from the embeddings. As a baseline, we evaluated the performance of out-of-the-box machine learning models: random forests, K-Nearest Neighbours (KNN), Multilayer Perceptron (MLP), Support Vector Machines (SVM), LGBM and XGBoost [47], [48]. Since we assumed that the robot would not get better over time, we did not need to resort to recurrent architectures such as [34], [49]. We found that most approaches suffer heavily from the cold start problem, *i. e.* an initial lack of data when only a few interactions have taken place [50]. To mitigate this and prevent the models from over-generalising

the initial observations, we also considered their performance with an ϵ-greedy condition.

To ensure our evaluation captured the models' ability to learn different proficiency profiles, we constructed a dataset using the Gazebo simulator [51]. This simulation replicated a robot's efforts to complete the pick and place task while contending with random failures in object detection, motion planning, and execution, each occurring at different frequencies. We established 10 distinct proficiency profiles, each executing the task 100 times. The objects in the environment were randomised every time, among a set of 21 possible objects. Furthermore, we conducted the same task in a real-world setting, collecting an eleventh dataset of 100 samples. To evaluate our predictors, we focused on two key criteria: the volume of data required by the model and the quality of its predictions. To fit these dual objectives, we followed the protocol detailed in Algorithm 1. The use of the Area Under the Receiving Operator Characteristic (AUROC) enabled us to capture both the overall prediction quality ("height" of the curve at each step) and learning speed (how fast it rises).

Algorithm 1 Evaluation of the self-confidence models

Require: (ζ, κ), the model to evaluate, $D = \{D_1, ..., D_{11}\}$ all datasets
acc $\leftarrow$ [ ]
for $D_i \in D$ **do**
  $D_i \leftarrow \zeta(D_i)$ ▷ *Text to embedding*
  $D_i^{train}, D_i^{test} \leftarrow D_i[0:80], D_i[80:100]$
  for $n_{train} \in [10:100:10]$ **do**
    $\kappa \leftarrow$ create_and_train$(D_i^{train}[:n_{train}])$
    acc $\leftarrow$ acc + eval_and_record(κ, D_i^{test})
  end for
end for ▷ acc *shape is* $(11, 10)$
acc $\leftarrow$ acc.mean() ▷ acc *shape is* $(1, 10)$
perf $\leftarrow AUROC(\text{acc})/90$ ▷ Normalise to $[0, 1]$

B. Experiment 2: Instructions following

Using semantic embeddings for action representations also offers a relatively simple solution for instructions following tasks. By projecting affordances descriptions and instructions in the same latent space, we can use a similarity metric to measure the relevance of our affordances. To evaluate SCONE's performance in this aspect, we used text recipes and ingredient lists from the Tasty V2 dataset to simulate kitchens with unconstrained agent capabilities [52]. Each run defined a kitchen as a space with randomly selected locations, such as cupboards or shelves, and spawned relevant and irrelevant ingredients based on the recipe. We used a PDDL-based agent to simulate the human's actions. This allowed us to generate 135 states, each made of an instruction i_s and a set of available affordances $\boldsymbol{\alpha}$. We then manually labelled 651 relevant actions in our 135 states, resulting in 135 triplets of (1) an instruction, (2) a set of relevant actions and (3) a set of irrelevant actions. 80% of those were used to train the custom models described later, and the remaining were used as a test set for all experiments. To measure success, we had the model predict the most relevant action in a labelled state until no more were available in that state. We then used these results to calculate accuracy as a ratio of relevant actions and averaged it over all the states.

In addition to the embeddings described in Section IV-A, we also used the training set to refine the MPNet embeddings. We added a MLP with dropout to the pre-trained MPNet and trained using a triplet loss based on the cosine distance in PyTorch [53]. We used the $651 * 0.8 = 520$ labelled (i_s, α^{rel}) text pairs as anchor and positive examples and used 10 random actions as negative examples for each pair, resulting in a dataset of 5200 samples. We tested different dimensions for our final embedding space and found the best results for a size of 128. Finally, we also considered two baselines specific to instructions following. Our first baseline was a PDDL planner, in which a Natural Language Processing (NLP)-based module was used to translate each instruction into a PDDL goal state. At each time step, the remaining recipe steps defined various PDDL goal states. The robot generated plans for each and picked one action from the set of initial actions of all generated plans. We implemented this baseline using Planutils [54] and the LAMA planner [55]. Our second baseline was an adjusted version of TICC-POMDP [6]; at each time step, the robot ran a Monte Carlo Tree Search (MCTS) which included its own proficiency estimate to find the best action. It updated its belief over this proficiency estimate at every time step.

C. Experiment 3: Effect on human trust

Finally, we ran a pilot study to evaluate the impact of self-confidence awareness on human trust. We considered a human and a robot standing on each side of a table, following a recipe to make a salad. The robot got the ingredients from a shelf and put them on the table, while the human used them to make the salad. SCONE was used to select the robot's actions based on the given recipe and objects perceived in the environment. An illustration of the setup is given in Figure 1, and more details are available in [56]. On the table, an iPad application served as a user interface, showing the recipe and the robot's current action. It also allowed the robot to ask for human help when a failure was predicted (in the session where self-confidence was used) or after 3 failures on the same affordance. Finally, it provided an option for the human to take over any action, used to measure an intervention ratio. We ran a within-participants experiment where each participant cooks twice with the robot: once with the self-confidence module active and once without. Our hypotheses were:

H1: Participants interacting with a self-confidence-aware robot will report a higher trust level.

H2: Participants interacting with a self-confidence-aware robot will have a lower interventions count.

We used two groups to account for familiarity with the setup, each encountering a different policy first. We collected age, gender and dietary restrictions using questionnaires from [57] before cooking sessions and assessed trust with questionnaires from both [57] and [58] after each session. To summarise, our independent variable is the robot's policy,

and the dependent variables are the number of interventions and reported trust levels from the various questionnaires. The study received approval from Imperial College's research ethics committee.

After both sessions, each participant gave Likert-type answers from the MDMT questionnaire's performance scale in its original form [58] and Likert-like from the TPS-HRI questionnaire [57]; see [59] for the distinction between Likert-type and Likert-like. This type of approach often results in longitudinal data, *i. e.* data with distributions that do not fit the statistical assumption of parametric hypothesis testing [59], [60]. A common alternative is to use non-parametric testing, which relaxes some of the distributional assumptions. However, this relaxation comes with a cost in terms of statistical power; for that reason, we preferred to use Bayesian data analysis to analyse the reported trust [61]. For comparability purposes, we also provide scoring results as described in both of the questionnaire's original papers. We do this while acknowledging that averaging ordinal data is controversial in statistical terms, as the 'distance' between every value of the scale is subjective [59]. Equation (3) shows a summary of the statistical model used for the questionnaires; both the MDMT and TPS-HRI models follow this structure. Equation (4) is the model used for the intervention count (some priors are identical to those in Equation (3) and not repeated). We invite readers to explore [61] for more details on Bayesian Statistical Modelling, and [60] for the usefulness of Cumulative Link Mixture Models (CLMMs) to model Likert scales.

$$
\begin{aligned}
R_i &\sim \text{Ordered-logit}(\phi_i, \alpha),\\
\phi_i &= \alpha_q[Q_i] + \alpha_p[P_i] + \alpha_g[G_i] + \alpha_c[C_i],\\
\alpha &\sim \mathcal{N}(0,1), \qquad \alpha_q[Q_i] \sim \mathcal{N}(\tilde{\alpha}, \sigma_q),\\
\tilde{\alpha} &\sim \mathcal{N}(0,1.5), \qquad \sigma_q \sim \text{Half-Normal}(0,1),\\
\alpha_p[P_i] &\sim \mathcal{N}(0,\sigma_p), \qquad \sigma_p \sim \text{Half-Normal}(0,1),\\
\alpha_g[G_i] &\sim \mathcal{N}(0,1), \qquad \alpha_c[C_i] \sim \mathcal{N}(0,1).
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
N_i &\sim \text{Poisson}(\lambda_i),\\
\log \lambda_i &= \alpha_p[P_i] + \alpha_g[G_i] + \alpha_c[C_i].
\end{aligned}
\tag{4}
$$

In Equations (3) and (4), the participant ID P_i is in $[\![1, 31]\!]$, G_i is 1 or 2 depending on whether this is the first or second cooking session, C_i is 1 or 2 depending on whether the self-confidence module was activated for that session, the question ID Q_i is in $[\![1, 8]\!]$ for the MDMT-performance questionnaire and in $[\![1, 14]\!]$ for TPS-HRI, and N_i is the number of interventions from the user. This model was fit using R and the Rethinking package's `ulam` function with 6 Markov chains of 500 warmup and 500 sampling iterations each [61]. This allowed us to sample the posterior distribution and compute the contrast, *i. e.* the distribution of the difference between the groups. The contrasts contain all of the information and uncertainty uncovered by our research and are a usual output of Bayesian Analysis. However, in an attempt to make a dichotomous decision and express our results in a fashion similar to the frequentist notion of null-hypothesis significance testing, we also report the overlap between the 95% Highest Density Interval (HDI) and the Region of Practical Equivalence (ROPE) [62].

	SC		IF
	LGBM	ϵ-KNN	
1HE	0.56	0.65	N/A
GloVe	0.61	0.69	0.69
BGE	**0.74**	0.79	0.75
MPNet	**0.74**	**0.81**	0.74
MPNet ft*	0.73	0.77	0.81
PDDL	N/A	N/A	**1.00**
Trust-TICC	N/A	N/A	0.80

*ft denotes further training on our custom dataset

TABLE I: **Self-confidence (SC)** and **Instructions Following (IF)** results. Metrics are in $[0, 1]$ and explained in section IV. LLM embeddings improve overall performance in both tasks (see subsection V-A).

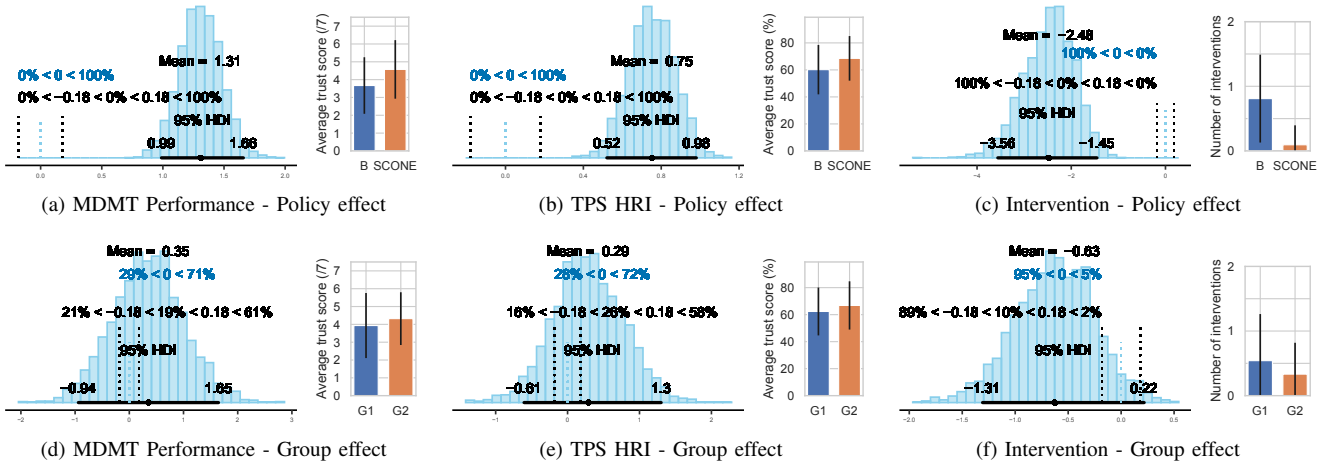
V. Results

A. Embeddings Performance

Since the embeddings we use are the same in Section IV-A and Section IV-B, we report the results together in Table I.

For the self-confidence prediction problem, we report the results from the best model candidates with and without the ϵ-greedy option. The best performance was overall obtained using transformers embeddings, with a slight advantage for MPNet (0.81) over BGE (0.79). The additional training slightly decreased the performance (0.77) but still left it significantly higher than other embeddings (0.69). Regarding the predictors, the use of the ϵ-greedy approach consistently improved the performance. Interestingly, LGBM was the best performer without it (0.74), but KNN was better with it (0.81). We believe this to be caused by KNN's lower bias, which reduces its tendency to overfit. This should be considered while keeping in mind that our metric emphasises learning efficiency; results would likely lean more toward higher bias models if we focused on mid to long-term performance. Additionally, users are likely to have personally preferred trade-offs between the robot asking for help or failing too often; future work could thus explore the impact of differentiating false positives and false negatives.

Regarding the instructions following task, the best results were obtained using MPNet with additional training (81%). Interestingly, the additional training only provided a marginal improvement, and the off-the-shelf models already performed reasonably (74% for MPNet and 75% for BGE). Compared to other approaches, ours selected relevant actions as often as [6] (81% versus 80%). This was of course lower than the PDDL baseline, which only selects relevant actions by design, but does not consider potential failures. Furthermore, we noted a few behaviours that could be explored in future work. First, we observed that the presence of negations in the instructions could result in a false positive increase. This was not a problem in our kitchen setting since few instructions asked the agent NOT to do something, but in different contexts, an instruction like 'move past the kitchen and go to the bedroom' could easily lead to a wrong outcome.



(a) MDMT Performance - Policy effect (b) TPS HRI - Policy effect (c) Intervention - Policy effect

(d) MDMT Performance - Group effect (e) TPS HRI - Group effect (f) Intervention - Group effect

Fig. 3: **Policy (top)** and **group (bottom)** effects on the questionnaire answers and the intervention ratio. **Left (histograms)**: HDI and ROPE of the contrasts. Limits are given by the low black horizontal line (HDI) and the dashed vertical lines (ROPE). Blue text shows if the effect is more positive or negative; black text shows how much of the contrast falls inside the ROPE. **Right (candle plots)**: average and standard deviation of the Likert scores (see Section IV-C) and number of interventions. **Figures (a), (b)**: HDI and ROPE are mutually exclusive, and the offset is positive. $H1$ is accepted, self-confidence awareness increases reported trust. **Figure (c)**: HDI and ROPE are mutually exclusive, and the offset is negative. $H2$ is accepted, self-confidence awareness reduces interventions. **Figures (d), (e), (f)**: the HDI and ROPE fully overlap in all group contrasts; the order in which a participant is exposed to the policies does not mitigate the effect (see Section V-B).

Second, we used a learnt threshold on the similarity metric as a stopping criterion; future work could investigate how this could be generalised to various contexts.

B. Impact on trust

We ran the experiments with 31 participants, with an average total time of 48 minutes. 55% of the participants were between in the [21-30] age range, 25% in [31-40], 13% in [41-50] and 7% in [51-60]. 52% identified as male and 48% as female. They reported various degrees of familiarity with automation. One question from the MDMT was skipped by a single participant and was dropped. As explained in Section IV-C, we used the overlap between the 95% HDI and the ROPE to decide whether or not to reject our hypotheses. Based on [62], we set our ROPE to $[-\frac{\epsilon}{2}\frac{\pi}{3}, \frac{\epsilon}{2}\frac{\pi}{3}]$ where ϵ is the effect size. Following [63], we used $\epsilon = 0.2$, resulting in a ROPE range of $[-0.18, 0.18]$. If the HDI fell fully inside the ROPE, we rejected the hypothesis as no significant difference was observed between the groups. If it fell fully outside, we accepted it. If there was only a partial overlap, we could not make a decision. Figure 3 plots the HDI and ROPE of the contrasts for the TPS-HRI, MDMT-performance and intervention count. The self-confidence-awareness provided by SCONE had a positive impact on reported performance-based trust, and reduced the number of interventions; $H1$ was accepted for both scales, and $H2$ was also accepted. Since this is a within-participant study, we are also interested in the impact of the order in which participants encountered the policies on their reported and behavioural trust. To answer this question, we use our Bayesian model to consider the contrast between both groups. As shown on Figure 3 where the HDI and ROPE fully overlap in all group contrasts, the participant's group does not affect their reported trust or number of interventions.

VI. CONCLUSION

In this paper, we explored the relevance of semantic action embeddings for assistive robots. We presented SCONE, a framework leveraging these embeddings to both learn self-confidence from experience and follow simple instructions. We showed that SCONE improved self-confidence accuracy and learning speed, while ensuring suitable action relevance. We also ran a pilot study with a realistic assistive cooking scenario, in which self-confidence awareness had a positive effect on capability-based human trust. While an effect of statistical significance was found, this study could be extended in several ways. The effect could be examined with a larger variety of ages, genders and familiarity with automation levels, as those have been found to be significant trust antecedents. Personality and culture have also been identified as important trust factors and could be considered [7]. Now that this pilot study confirmed the presence of an effect, future work can be done on quantifying its magnitude and further investigating it. Furthermore, our approach focused on the first step towards building trust in HRI; future research could then explore the human perception of robot proficiency and how to integrate it into planning. Such an extension could allow the robot to actively calibrate the human's trust based on their capabilities.

ACKNOWLEDGEMENTS

The authors would like their colleagues from the Personal Robotics Laboratory for their help proofreading this manuscript, and especially Rodrigo Chacón Quesada for his help with the Bayesian analysis.

REFERENCES

[1] G. M. Alarcon, A. L. Baker, M. J. Barnes, S. Berman, J. P. Bliss, J. M. Bradshaw *et al.*, *Trust in Human-Robot Interaction*, 1st ed., N. C. S. and . L. J. B, Eds. Academic Press, 11 2020. [Online]. Available: https://shop.elsevier.com/books/trust-in-human-robot-interaction/nam/978-0-12-819472-0

[2] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon *et al.*, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3–12. [Online]. Available: https://doi.org/10.1145/3319502.3374793

[3] A. Norton, H. Admoni, J. Crandall, T. Fitzgerald, A. Gautam, M. Goodrich *et al.*, "Metrics for robot proficiency self-assessment and communication of proficiency in human-robot teams," *J. Hum.-Robot Interact.*, vol. 11, no. 3, jul 2022. [Online]. Available: https://doi.org/10.1145/3522579

[4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 7 2023. [Online]. Available: https://arxiv.org/abs/2307.15818v1

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[6] J. Lee, J. Fong, B. C. Kok, and H. Soh, "Getting to know one another: Calibrating intent, capabilities and trust for human-robot collaboration," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 6296–6303.

[7] P. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, "Evolving trust in robots: specification through sequential and comparative meta-analyses," *Human factors*, vol. 63, no. 7, pp. 1196–1229, 2021.

[8] K. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors The Journal of the Human Factors and Ergonomics Society*, vol. 57, pp. 407–434, 05 2015.

[9] C. Textor, R. Zhang, J. Lopez, B. G. Schelble, N. J. McNeese, G. Freeman *et al.*, "Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach," *Journal of Cognitive Engineering and Decision Making*, vol. 16, no. 4, pp. 252–281, 2022. [Online]. Available: https://doi.org/10.1177/15553434221113964

[10] N. Conlon, D. Szafir, and N. Ahmed, ""i'm confident this will end poorly": Robot proficiency self-assessment in human-robot teaming," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2127–2134.

[11] X. Cao, A. Gautam, T. Whiting, S. Smith, M. A. Goodrich, and J. W. Crandall, "Robot proficiency self-assessment using assumption-alignment tracking," *IEEE Transactions on Robotics*, 8 2023.

[12] T. Frasca and M. Scheutz, "A framework for robot self-assessment of expected task performance," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 523–12 530, 10 2022.

[13] B. Israelsen, N. Ahmed, E. Frew, D. Lawrence, and B. Argrow, "Machine self-confidence in autonomous systems via meta-analysis of decision processes," *Advances in Intelligent Systems and Computing*, vol. 965, pp. 213–223, 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-20454-9_21

[14] A. S. Bauer, P. Schmaus, F. Stulp, and D. Leidner, "Probabilistic effect prediction through semantic augmentation and physical simulation," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 9278–9284, 5 2020.

[15] P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev, and C. Schmid, "Instruction-driven history-aware policies for robotic manipulations," 2022. [Online]. Available: https://arxiv.org/abs/2209.04899

[16] H. Liu, L. Lee, K. Lee, and P. Abbeel, "Instruction-following agents with jointly pre-trained vision-language models," *arXiv preprint arXiv:2210.13431*, 2022.

[17] D. P. Losey, K. Srinivasan, A. Mandlekar, A. Garg, and D. Sadigh, "Controlling assistive robots with learned latent actions," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 378–384, 5 2020.

[18] S. Sharma, J. Gupta, S. Tuli, R. Paul, and Mausam, "Goalnet: Inferring conjunctive goal predicates from human plan demonstrations for robot instruction following," 2022. [Online]. Available: https://arxiv.org/abs/2205.07081

[19] M. Murray and M. Cakmak, "Following natural language instructions for household tasks with landmark guided search and reinforced pose adjustment," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6870–6877, 2022.

[20] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay *et al.*, "Progprompt: Generating situated robot task plans using large language models," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11 523–11 530, 5 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10161317/

[21] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," 12 2022. [Online]. Available: https://arxiv.org/abs/2212.04088v3

[22] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter *et al.*, "Code as policies: Language model programs for embodied control," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 5 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10160591/

[23] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," pp. 9118–9147, 6 2022. [Online]. Available: https://proceedings.mlr.press/v162/huang22a.html

[24] B. F. Malle and D. Ullman, "Chapter 1 - a multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds. Academic Press, 2021, pp. 3–25. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128194720000010

[25] T. Sassmannshausen, P. Burggräf, M. Hassenzahl, and J. Wagner, "Human trust in otherware - a systematic literature review bringing all antecedents together," *Ergonomics*, vol. 66, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36062352/

[26] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995. [Online]. Available: https://doi.org/10.5465/amr.1995.9508080335

[27] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011, pMID: 22046724. [Online]. Available: https://doi.org/10.1177/0018720811417254

[28] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004, pMID: 15151155. [Online]. Available: https://doi.org/10.1518/hfes.46.1.50_30392

[29] G. M. Alarcon, A. L. Baker, M. J. Barnes, S. Berman, J. P. Bliss, J. M. Bradshaw *et al.*, "Contributors," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds. Academic Press, 2021, pp. xiii–xvi. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128194720099925

[30] Y. Xie, I. P. Bodala, D. C. Ong, D. Hsu, and H. Soh, "Robot capability and intention in trust-based decisions across tasks," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 39–47.

[31] P. Kulms and S. Kopp, "More human-likeness, more trust? the effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation," in *Proceedings of Mensch Und Computer 2019*, ser. MuC'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 31–42. [Online]. Available: https://doi.org/10.1145/3340764.3340793

[32] N. Lingg and Y. Demiris, "Beyond self-report: A continuous trust measurement device for hri," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. Institute of Electrical and Electronics Engineers (IEEE), 11 2023, pp. 2220–2225. [Online]. Available: https://doi.org/10.1109/RO-MAN57019.2023.10309660

[33] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE robotics and automation letters*, vol. 6, pp. 5913–5920, 2021.

[34] H. Soh, Y. Xie, M. Chen, and D. Hsu, "Multi-task trust transfer for human–robot interaction," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 233–249, 2020.

[35] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human-robot collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot*

Interaction, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 307–315. [Online]. Available: https://doi.org/10.1145/3171221.3171264

[36] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 221–228.

[37] S. Zörner, E. Arts, B. Vasiljevic, A. Srivastava, F. Schmalzl, G. Mir *et al.*, "An immersive investment game to study human-robot trust," *Frontiers in Robotics and AI*, vol. 8, p. 139, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.644529

[38] N. Lingg and Y. Demiris, "Building trust in assistive robotics: Insights from a real-world mobile navigation experiment," *ACM International Conference Proceeding Series*, 7 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3597512.3597519

[39] J. J. Gibson, "The ecological approach to the visual perception of pictures," *Leonardo*, vol. 11, no. 3, pp. 227–235, 1978.

[40] R. Chacón-Quesada and Y. Demiris, "Proactive robot assistance: Affordance-aware augmented reality user interfaces," *IEEE Robotics & Automation Magazine*, vol. 29, no. 1, pp. 22–34, 2022.

[41] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5913–5920, 2021.

[42] J. Mi, S. Tang, Z. Deng, M. Goerner, and J. Zhang, "Object affordance based multimodal fusion for natural human-robot interaction," *Cognitive Systems Research*, vol. 54, pp. 128–137, 5 2019.

[43] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," 2020. [Online]. Available: https://arxiv.org/abs/2004.09297

[44] BAAI, "Flagembedding," https://github.com/FlagOpen/FlagEmbedding, 2023.

[45] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[46] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[47] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[48] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[49] A. Cully and Y. Demiris, "Online knowledge level tracking with data-driven student models and collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2000–2013, 2020.

[50] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, "A collaborative filtering approach to mitigate the new user cold start problem," *Knowledge-based systems*, vol. 26, pp. 225–238, 2012.

[51] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2149–2154, 2004.

[52] F. Sener, R. Saraf, and A. Yao, "Transferring knowledge from text to video: Zero-shot anticipation for procedural actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2022.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[54] C. Muise, F. Pommerening, J. Seipp, and M. Katz, "Planutils: Bringing planning to the masses," in *32nd International Conference on Automated Planning and Scheduling, System Demonstrations and Exhibits*, 2022.

[55] S. Richter and M. Westphal, "The lama planner: Guiding cost-based anytime planning with landmarks," *Journal of Artificial Intelligence Research*, vol. 39, pp. 127–177, 2010.

[56] C. Goubard and Y. Demiris, "Cooking up trust: Eye gaze and posture for trust-aware action selection in human-robot collaboration," *ACM International Conference Proceeding Series*, p. 5, 7 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3597512.3597518

[57] K. E. Schaefer, *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*. Boston, MA: Springer US, 2016, pp. 191–218. [Online]. Available: https://doi.org/10.1007/978-1-4899-7668-0_10

[58] B. F. Malle and D. Ullman, "Measuring human-robot trust with the mdmt (multi-dimensional measure of trust)," 11 2023. [Online]. Available: https://arxiv.org/abs/2311.14887v1

[59] M. Schrum, M. Ghuy, E. Hedlund-botti, M. Natarajan, M. Johnson, and M. Gombolay, "Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices," *J. Hum.-Robot Interact.*, vol. 12, no. 3, apr 2023. [Online]. Available: https://doi.org/10.1145/3572784

[60] J. E. Taylor, G. A. Rousselet, C. Scheepers, and S. C. Sereno, "Rating norms should be calculated from cumulative link mixed effects models," *Behavior Research Methods*, vol. 55, no. 5, pp. 2175–2196, Aug 2023. [Online]. Available: https://doi.org/10.3758/s13428-022-01814-7

[61] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[62] J. K. Kruschke and T. M. Liddell, "The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective," *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 178–206, Feb 2018. [Online]. Available: https://doi.org/10.3758/s13423-016-1221-4

[63] J. Kruschke, "Doing bayesian data analysis: A tutorial with r," *JAGS, and Stan*, vol. 2, 2014.