Reinforcement Learning-Based Intent-Action Dialogue Manager

Anna Manaseryan
Boise State University
annamanaseryan@u.boisestate.edu

Abstract

Developing adaptive dialogue systems that can accurately map natural language commands to robotic actions is an important area in Human-Robot Interaction (HRI). This paper presents a two-stage learning framework designed to create a more robust intent-action dialogue manager. Our methodology first establishes a foundational policy by performing Supervised Fine-Tuning. Furthermore, this policy is refined using Reinforcement Learning (RL) with the Proximal Policy Optimization (PPO) algorithm. The evaluation demonstrates that the RL-refined policy achieves a significantly higher mean reward and a more consistent reward distribution compared to the initial supervised baseline. These findings indicate that our approach is an effective method for enhancing the accuracy and reliability of dialogue systems for robotic control, providing a good foundation for future applications and human evaluation in real-world environments.

1 Introduction

A main goal in artificial intelligence is to build systems that can think and act independently, similar to how humans do. In the field of machine learning, **reinforcement learning (RL)** is especially important for creating self-operating robots because it closely replicates how the human brain learns (Singh et al., 2022). For a robot or smart system to work properly, it must easily understand a person's broad and long-term commands and turn them into a series of smaller actions. This process is a key part of systems designed to help with tasks through conversation, as it connects what a person says to what the system can do (Feng et al., 2025). This is a challenging task because a system's view of its environment and its grasp of what a person wants are often incomplete or uncertain (Hatanaka et al., 2023).

Traditional methods of control often face trouble solving the complex and uncertain problems that are part of human-robot interaction (Singh et al., 2022). Even current approaches that use supervised learning, including Large Language Models (LLMs), can perform much worse when they face new or unexpected requests from a user (Feng et al., 2025). This shows the need for a more flexible and adaptable method. Standard supervised fine-tuning often does not generalize well, making it hard for systems to handle new tasks without constantly retrained (Feng et al., 2025). Reinforcement learning offers a powerful solution that gives systems a way to learn and adapt in changing and uncertain environments (Singh, 2022). RL has proven to be useful for helping large models generalize better (Feng et al., 2025), allowing a system to learn the best course of action through trial-and-error and feedback in the form of rewards (Švaco et al., 2018).

This paper looks at how Reinforcement Learning can be used to solve the key problems of uncertainty and adaptation. By treating the conversation process as an RL problem, we can build a system that learns the best strategy for connecting a user's request to the robot's actions in order to get the best result over time. In our setup, the part that manages the conversation acts as the agent, what the user says represents the current state, and the task the robot chooses is the action. The main goal of our research is to build and test a conversation manager that uses RL to continuously improve the connection of user requests to robot actions. Improving how well these models can understand new and different requests is still a very important task in this field (Feng et al., 2025). Therefore, we want to answer the following research question: How much can a system using reinforcement learning, which learns from user feedback, improve the accuracy and flexibility of turning spoken commands into robot actions? Our guess is that a conversation manager based on RL, will do better than models that don't learn from experience. What this work could

add is a stronger and more natural way for humans and robots to communicate, allowing robots to adjust to different users and new situations without needing much retraining.

2 Background

The pursuit of adaptive dialogue systems has seen significant progress with the adoption of Reinforcement Learning (RL), moving beyond static rule-based agents. Early applications of RL successfully optimized dialogue efficiency and task completion rates in virtual agents, such as those for customer service or reservation systems (Li et al., 2016). These foundational studies demonstrated that an agent could learn an effective communication policy through trial and error. However, the simple reward functions used in these systems often do not account for the complexities of real-world interaction.

Addressing the inherent ambiguity in human language has focused on equipping agents with the ability to manage their own uncertainty. Building on work in active learning, models have been developed that learn a query policy, allowing them to recognize when they lack sufficient confidence and should ask the user for clarification (Hatanaka et al., 2023). This represents a critical step toward more robust communication. Our work builds upon this concept by designing a framework intended to be applicable in physical contexts.

Another advancement in RL has been the move toward more sophisticated reward mechanisms to guide agent learning. The challenge of designing effective rewards is well documented, and Serban et al. (2017) have begun to leverage large-scale models to provide dense, context-aware feedback. Our research proposes a framework at the confluence of these developments. The objective is to design a dialogue manager that integrates an uncertainty-aware query policy with a model-based reward structure, with the objective of application in the domain of human-robot interaction. This framework is intended to provide a foundation for a system in which a robot could learn an optimal action strategy while also managing ambiguity.

3 Data

For this research, we built a dataset specifically to train and test our model's ability to match a person's spoken command to one of five predefined robot actions which are speak, detect objects, move, pick up, and express emotion. To build this dataset, we used a mix of existing and newly created data. For the speak action, we used data from the QA-Assistant-2 dataset (Mihaiii) designed for dialogue assistants, as its question format was a good fit for handling informational requests that require a spoken response. For the other four actions, we generated our own custom data providing language variety and creating a wide range of examples.

4 Model & Approach

To address the challenge of creating an intent-action dialogue manager, we propose a learning framework that uses reinforcement learning to refine a foundational policy. Our approach is designed to first establish an understanding of the task through a fine-tuned model and then iteratively improve its decisions through automated feedback. This methodology allows the system to learn a mapping from natural language commands to executable robotic actions while continuously improving its precision and adaptability.

This framework is implemented using three main components: a policy model, a reward model, and an optimization algorithm. The **policy model** is the core of our dialogue manager, responsible for selecting the robot action. We use a fine-tuned model, which takes a user's command as input and outputs a predicted action label. The **reward model**, used during the reinforcement learning stage, provides feedback. For this, we use a larger model that takes the user's command and the selected action as input and produces a score indicating the action's quality. This feedback is then used by the **Proximal Policy Optimization (PPO)** algorithm to improve the policy model. This setup allows the policy model to be trained using the feedback from the larger reward model. Further details on this process will be provided in the following sections.

5 Task & Procedure

In this section, we present a dialogue manager that can learn to understand a user's intent from a natural language question and then select the correct robotic action. This method will allow the model to improve itself through a feedback loop based on reinforcement learning. The overall process has two main stages: initial policy training and policy refinement.

5.1 Initial Policy Training

First, we establish a foundational policy model that has a basic understanding of how to map questions to actions. To do this, we fine-tune a Flan-T5 model on a custom dataset of *(question, action)* pairs. This training process teaches the model the basic connections between language and the available robot actions. The result of this stage is an initial policy model. This model serves as a strong starting point, but its knowledge is limited to what it saw in the initial dataset, and it cannot adapt to new or different user requests.

5.2 Policy Refinement with Reinforcement Learning

The second stage makes this initial policy more adaptive and precise. To achieve this, we use a reinforcement learning framework that requires a reward model to provide feedback. First, we used a LLaMA3-70b-Instruct model as judge to score (query, action) pairs, creating a reward dataset. Then, we trained a smaller, more efficient DistilBERT model on this dataset. This DistilBERT model learned to predict the reward scores that the larger LLaMA model would give, and it became the official reward model for our training process.

With this reward model in place, we refine our Flan-T5 policy using the Proximal Policy Optimization algorithm. In this training loop, the Flan-T5 policy model receives a query and generates an action. The DistilBERT reward model then scores that action, providing a reward signal. The PPO algorithm uses this reward to update and improve the Flan-T5 model, encouraging it to make better decisions in the future. This process of using feedback from an automated reward system is a key technique for building adaptive dialogue systems (Li et al., 2016).

5.3 Evaluation

The primary goal of our evaluation was to measure the improvement gained from the reinforcement learning stage. While a full human evaluation in a physical robotic environment is planned as a future step, this initial analysis compares the performance of the final policy against the baseline policy using automated metrics. The evaluation was performed on a test that was set of user queries not seen during training.

Metrics To analyze performance, we used two key metrics based on the scores from the reward model:

- **Mean Reward** This provides a summary of a policy's overall performance by calculating the average score across the test set.
- **Reward Distribution** This offers a more detailed view by showing the frequency of high-quality versus low-quality actions, which helps assess model consistency.

Results The analysis shows improvements from the reinforcement learning phase. The final policy achieved significantly higher mean reward than the baseline policy (Figure 1), indicating an overall improvement in action quality. Furthermore, the reward distribution in Figure 2 reveals that the PPO model became more consistent, with its scores concentrated in the higher range and a reduction in low-quality predictions. These metrics suggest that our framework produces a more accurate and reliable policy, providing a foundation for future settings.

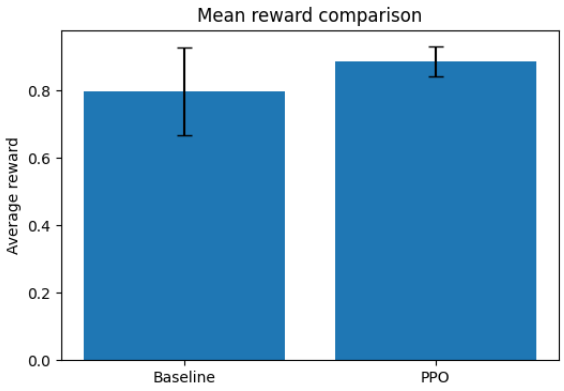

Figure 1: A comparison of the reward scores for the baseline supervised model and the final PPO-tuned model.

6 Conclusion

In this work, we addressed creating a dialogue manager that is capable of mapping natural language commands to robotic actions. We implemented a two-stage learning framework that first establishes a baseline policy with supervised fine-tuning and then refines it using reinforcement learning with Proximal Policy Optimization (PPO).

These results demonstrate that this reinforcement learning phase provides significant benefits. The final, refined policy showed improvements in

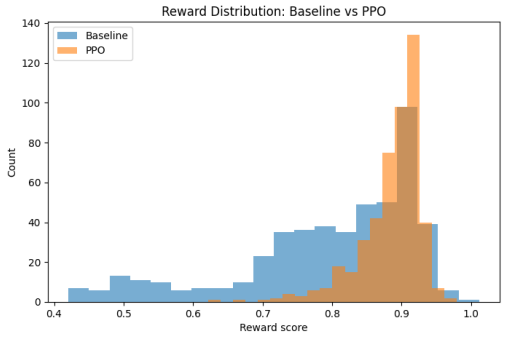

Figure 2: The distribution of reward scores for the baseline and PPO-tuned models on the test set. The PPO model's distribution shows a significant rightward shift and a more concentrated peak around a high reward score, indicating a consistent improvement in the quality of its generated actions compared to the more varied performance of the baseline model.

both average performance and consistency when compared to the initial supervised model, according to our metrics. This suggests that our framework is effective in producing a more accurate and reliable policy for understanding user intent.

This research highlights the potential of combining supervised methods with reinforcement learning to build more robust human-robot communication systems. The next step for this work is to deploy our trained policy in a physical robotic environment. In future work we plan to explore applying this framework to more complex and multi-turn dialogue tasks.

References

Zihao Feng, Xiaoxue Wang, Ziwei Bai, Donghang Su, Bowen Wu, Qun Yu, and Baoxun Wang. 2025. Improving generalization in intent detection: Grpo with reward-based curriculum sampling. *arXiv preprint arXiv:2504.13592*.

Wataru Hatanaka, Ryota Yamashina, and Takamitsu Matsubara. 2023. Reinforcement learning of action and query policies with ltl instructions under uncertain event detector. *IEEE Robotics and Automation Letters*, 8(11):7010–7017.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Mihaiii. qa-assistant-2. https://huggingface.co/datasets/Mihaiii/qa-assistant-2. Dataset hosted on Hugging Face.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. 2022. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990.

Nikhil Singh. 2022. niksss at qur'an QA 2022: A heavily optimized BERT based model for answering questions from the holy qu'ran. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 126–129, Marseille, France. European Language Resources Association.

Marko Švaco, Bojan Jerbić, Mateo Polančec, and Filip Šuligoj. 2018. A reinforcement learning based algorithm for robot action planning. In *International Conference on Robotics in Alpe-Adria Danube Region*, pages 493–503. Springer.