2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Interactive Reinforcement Learning from Natural Language Feedback

Imene Tarakli[1], Samuele Vinanzi[1], Alessandro Di Nuovo[1]

***Abstract*— Large Language Models (LLMs) are increasingly influential in advancing robotics. This paper introduces ECLAIR (Evaluative Corrective Guidance Language as Reinforcement), a novel framework that leverages LLMs to interpret and incorporate diverse natural language feedback into robotic learning. ECLAIR unifies various forms of human advice into actionable insights within a Reinforcement Learning context, enabling more efficient robot instruction. Experiments with real-world users demonstrate that ECLAIR accelerates the robot's learning process, aligning its policy closer to optimal from the outset and reducing the need for extensive human intervention. Additionally, ECLAIR effectively integrates multiple types of advice and adapts well to prompt modifications. It also supports multilingual instruction, broadening its applicability and fostering more inclusive human-robot interactions. Project website: `https://sites.google.com/view/eclairiros`**

I. INTRODUCTION

Reinforcement Learning (RL) has significantly advanced the field of robotics by allowing robots to learn diverse tasks ranging from navigation [1] to complex control policies [2]. Within this framework, robots autonomously acquire optimal behaviours through trial and error, using a predefined reward function to assess their performance and update their policy [3]. Despite RL's success in many robotics applications, it faces notable challenges, especially in real-world scenarios. First, RL heavily relies on the reward function; designing an effective reward function that captures all the task dimensions without promoting unintended behaviours is complex, time-consuming, and requires a deep technical insight [4], [5]. Furthermore, while autonomous exploration is essential for discovering optimal policies, extensively exploring a large set of behaviours can slow policy convergence and potentially lead to hazardous situations that compromise robot safety [6].

To address the challenges of designing accurate reward functions, various studies have explored leveraging human expertise in the learning process of robots through Interactive RL. In this approach, individuals knowledgeable about the task, yet possibly lacking in technical skills, guide the robot's learning by providing advice. This advice enables the robot to adapt and learn from diverse human inputs, facilitating faster policy convergence and ensuring safer exploration [7].

Teaching a robot involves different teaching strategies with various teaching signals [8]. Evaluative feedback directly assesses the optimality of a robot's action; either through scalar or binary values or by comparing demonstrated trajectories. This approach enables robots to learn without a predefined reward function, leading to quicker policy convergence. Corrective feedback, on the other hand, allows humans to refine the robot's policy by specifying the optimal action, thereby narrowing the exploration space and accelerating convergence. Another form of human advice is guidance, where humans guide the learning of the robot by informing it about future aspects of the task. This allows the robots to quickly converge to the optimal policy by reducing the randomness of the exploration process.

All these types of teaching signals contribute to alleviating the limitations of autonomous learning by allowing a faster convergence and safer exploration for robots [8]. Nevertheless, integrating human advice into the learning process is not straightforward. It demands significant engineering efforts to ensure robots can interpret these signals correctly, often constraining how humans naturally teach. Evaluative feedback, for example, can be directly included in the RL framework as an evaluation function, it, however, does not cover the richness of human knowledge as it limits the information that can go through this channel, as the users can only comment on the optimality of past actions [9]. Corrective feedback and guidance, on the other hand, offer more control to humans by allowing them to direct the robot toward optimal states or actions. However, interpreting such teaching signals is more complex as it necessitates a sophisticated mapping between the advice and its corresponding actions or states. This often relies on heavily annotated data [10], [11], multiple task demonstrations [12], [13], or predefined performance metrics to ground the teaching signals effectively [14]. Moreover, although significant efforts have been deployed to combine different teaching methods, teaching methods have been mostly investigated individually. A unified formalism of all these teaching signals remains an active research question [8].

Large Language Models (LLMs), with their demonstrated proficiency in learning in context and capturing essential commonsense priors about human behaviour [15], present a novel approach to understanding and integrating human feedback into robotic systems. In this work, we leverage these capabilities to introduce the **E**valuative **C**orrective Guidance **LA**nguage as re**I**nfo**R**cement **(ECLAIR)** model. ECLAIR is designed to seamlessly integrate diverse types of feedback provided in natural language into a unified framework that shapes robot behaviours. We aim to reduce the advice interpretation limitations of previous models by using pre-trained LLMs to interpret various human advice and effectively translate it into actionable insights that inform and enhance the robot's learning process. Our contributions are as follows:

[1]Department of computing, Sheffield Hallam University, S1 1WB, Sheffield `i.tarakli@shu.ac.uk`

- We present ECLAIR, a pioneering model that leverages LLMs to unite evaluative, corrective, and guidance feedback within a single framework for interactive teaching.
- We empirically validate ECLAIR's effectiveness in accelerating robot learning and reducing the human training load in the process.
- We demonstrate that ECLAIR is robust against varying prompt structures and can be used across multiple languages.

II. RELATED WORK

Recent research has explored leveraging LLMs to enhance Reinforcement Learning performance. Various approaches include utilising LLMs to generate rewards by finetuning on extensive user data [16], [17], or employing in-context learning strategies with limited datasets [18], [19]. Other studies have enhanced RL by employing LLMs to guide the learning of intermediate tasks with language directives [20], [21] or by creating goals for an agent and defining rewards through the cosine similarity between the goal's description and the observation's caption [22]. To the best of our knowledge, no existing studies have investigated the integration of LLMs for directly interpreting human advice within Interactive RL frameworks.

III. PRELIMINARIES

Reinforcement Learning (RL) focuses on solving tasks represented as Markov Decision Processes (MDPs) [3]. Defined by the tuple $< S, A, T, R, \gamma >$, an MDP consists of S and A as the sets of states and actions, respectively. The transition function $T : S \times A \rightarrow S$ determines the probability of transitioning to a new state given a current state and action. The reward function $R : S \times A \rightarrow R$ outlines the reward for executing an action in a state, while $\gamma \in [0, 1]$ represents the discount factor, which influences the agent's sensitivity to future rewards.

The aim in RL is to identify policies $\pi : S \rightarrow A$ that optimise the total expected discounted rewards over time. The expected return from each state-action pair, referred to as the action-value, is denoted by $Q(s, a)$ and is defined as $Q(s, a) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s, a)\right]$. Optimal policies, symbolised as π^*, are those that maximise these returns, guiding the agent to achieve the highest rewards.

IV. THE ECLAIR FRAMEWORK

In this section, we present Evaluative Corrective Guidance LAnguage as reInfoRcement (ECLAIR), a RL framework that integrates different types of natural language feedback to interactively shape robots' behaviours. The model consists of two phases:

1) **Advice interpretation:** we leverage the use of LLMs to translate the spoken feedback into different value, specifically evaluative feedback, corrective feedback, and guidance for the next action.
2) **Advice shaping:** this consists of integrating the different types of feedback in the RL algorithm to update and refine the policy of the robot.

A. Advice interpretation

In this phase, natural language feedback from humans is interpreted and transformed into an internal representation that guides the robot's learning process. Pre-trained LLMs act as interpreters of human advice, converting spoken input into actionable feedback. This feedback is categorised into three types: evaluative, which provides a binary assessment of an action's correctness; corrective, which suggests alternative actions the robot should take; and guidance, which offers direction for future actions. To facilitate this process, we refine the LLM's system prompt by detailing the task, its objectives, and the action space within the Markov Decision Process (MDP). We clearly define each type of feedback and use few-shot prompting to ensure accurate interpretation. By providing examples of human advice alongside their corresponding outputs, we help the LLMs understand the interpretation task and adhere to the required format. The system prompt used is included in the appendix and is also available on the project page. This phase occurs after the robot completes an action. If the user provides advice, a prompt is generated that combines the robot's action with the human advice provided, as illustrated in Figure 1. This prompt is processed by the LLM, which outputs the interpreted feedback. A parser then converts this textual feedback into numerical values, ready for integration into the robot's learning process. If no advice is given, or if it does not cover all feedback types, a default 'None' value is assigned to the specific type of feedback, and the robot proceeds to the next action.

B. Advice shaping

In this phase, we integrate the interpreted human advice into the RL system to shape the learning process of the robots. For this, we take inspiration from existing shaping strategies and integrate the advice into the learning process differently for each type of feedback.

Prior studies have demonstrated that evaluative feedback is most effectively utilised as immediate information about the value of an action, positioning it as an ideal candidate for directly modifying the action-value function within the learning algorithm [23]. Similar to Knox et al.[24], we adopt the Q-augmentation method to integrate the evaluative feedback, denoted as h, into the learning process. For each observed state-action pair $< s, a >$, we update the Q-value as follows:

$$Q'(s, a) = Q(s, a) + \alpha * (h - Q(s, a)) \tag{1}$$

When the robot selects an incorrect action, the corrective feedback specifies the optimal action, denoted as a_c, that would have been more appropriate. For its integration into the learning process, we employ a methodology paralleling Q-augmentation; instead of altering the action-value of the observed state-action pair $< s, a >$, we modify the Q-value for the state and the optimal action $< s, a_c >$ that the robot should have selected. This method of adjustment is essentially equivalent to providing positive feedback for the

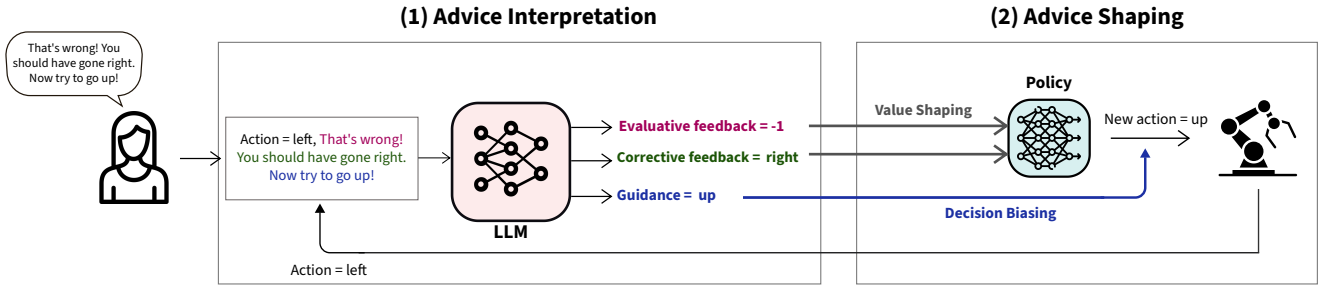


Fig. 1: Illustration of the framework. (1) The agent first learns a low-level policy with a myopic interactive RL. All trajectories of the interaction are stored in a buffer. (2) An offline inverse RL is then applied to the stored trajectories to recover a reward and a policy that better encodes the high-level information of the task.

corrected action a_c. The update is formulated as follows:

$$Q'(s, a_c) = Q(s, a_c) + \alpha(1 - Q(s, a_c)) \tag{2}$$

Guidance is provided to steer the exploration strategy toward the desired action of the user. Unlike evaluative and corrective feedback, which influence the learning algorithm's value estimations, guidance feedback directly modifies the policy's output at decision time. To incorporate this feedback into the learning process, we consider the decision-biasing strategy where the robot prioritises the action suggested by the user, denoted as a_g, at decision time. Specifically, the policy is adjusted such that for the next state s_{t+1}, the chosen action is explicitly set to a_g, formalised as $\pi(s_{t+1}) = a_g$.

V. METHODOLOGY

We structure our methodology to address the following research questions:

- Does ECLAIR enhance the process of learning from human advice?
- Is ECLAIR feedback efficient?
- What is the sensitivity of ECLAIR to variations in the prompt structure, and how does it impact the advice interpretation?

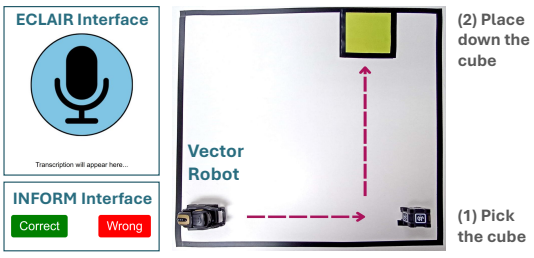


Fig. 2: Experimental setup: Vector robot, *pick and place* task and screenshots from the interfaces.

A. Experimental setup

To assess the performance of ECLAIR, we designed a robotics study where participants are asked to teach a robot a control task. Figure 2 illustrates the experimental setup.

Robotics platform. Vector.[1], a compact social robot is used in this study. Measuring a few inches, this robot is particularly suitable for tabletop interactions. Designed for dynamic interaction with its surroundings, Vector is equipped with navigation and manipulation abilities. Its wheels enable omnidirectional movement to navigate diverse surfaces. Additionally, Vector is equipped with a lift mechanism, allowing it to engage directly with objects, such as picking up or setting down its interactive cube.

System configuration. The setup consists of three elements – the robot, an Android tablet, and an Ubuntu machine. The Android tablet serves as the interactive interface through which humans can deliver advice. The Ubuntu machine runs the RL process and uses the Robot Operating System (ROS) to facilitate communication among all three elements.

LLMs and STT. To interpret the human advice, we use the gpt-3-turbo model [15] that was demonstrated to be efficient in generating and understanding human texts. Given that the input must be in a textual format, we utilise Whisper [25], a flexible and robust Speech To Text (STT) model, to transcribe spoken advice provided by the user.

RL algorithm. We use Q-learning [3] as the RL process of the study. The algorithm is suitable for discrete cases and often ensures convergence to an optimal policy. This allows us to thoroughly evaluate ECLAIR's performance as a comprehensive advice interpretation system.

B. Task domain

We consider a *pick and place* control task where the robot's objective is to start from a specified position, navigate towards a cube, pick it up, and subsequently place it in a designated location.

We model the environment as a grid consisting of 25 cells, determining the robot's precise cell location, L_{cell}, through its xy coordinates. The task state space is then defined by the tuple $S = (L_{cell}, \text{is\_picked})$ where L_{cell} represents the robot's current location, and is_picked is a binary value assessing whether the cube has been picked up, yielding a state space size of 50 states.

[1] https://ddlbots.com/products/vector-robot

Moreover, the robot is able to perform six primitive actions; the robot can go up, down, left, right, pick the cube, and put it down. Each episode begins with the robot at a location marked by an X and concludes either when the cube is placed down or after 15-time steps, whichever occurs first.

The objective for the robot is to accomplish the task with the minimum number of time steps. To evaluate the robot's performance, we use the negative of the episode's step count. Since prematurely dropping the cube can result in the episode concluding in fewer steps than the optimal path would require, we adjust this metric to account for such outcomes. Specifically, we impose a penalty of -15 if the cube is not picked up by the end of the episode and a -10 penalty if the cube is placed in a location other than the designated target spot.

C. Study design

To evaluate the performance of ECLAIR, we employ a within-subjects design wherein participants instruct Vector to perform a control task using two distinct teaching approaches: ECLAIR and TAMER [26]. TAMER, which relies exclusively on evaluative feedback, serves as the control condition in our study. Our selection of TAMER was motivated by its straightforward implementation and high replicability, providing a good baseline for comparison. This design enables us to study how multiple advice channels, as featured in ECLAIR, influence robot learning in contrast to TAMER's single-channel approach.

We recruited 12 participants ($M_{age} = 30$, 10 men, 2 women), from Sheffield Hallam University to take part in our study. Participants were introduced to Vector robot and briefed on the control task and teaching methods. To mitigate any potential order effects, we randomised the order of the two teaching methods.

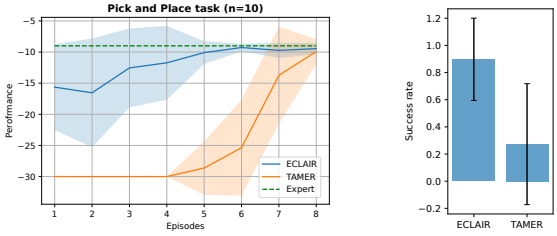
For each condition, participants were instructed to guide the robot through the task across 8 episodes. This specific number of episodes was strategically chosen to balance the overall study duration and accommodate the robot's battery life, and was based on preliminary simulations indicating that both teaching methods would achieve convergence under a rational teacher's guidance.

For ECLAIR, participants provide their advice through speech using an interface and the microphone of an Android tablet. In the TAMER condition, evaluative feedback is provided through a different interface featuring two buttons labelled "correct" and "wrong." This specific design for TAMER was chosen to strictly limit feedback to evaluative inputs, addressing findings from previous research where participants often repurposed evaluative channels to offer guidance [9].

Before the start of the teaching sessions, each participant underwent a trial episode in each condition to familiarise themselves with the task and interfaces. On average, the study duration was 1 hour and 30 minutes per participant, cumulatively yielding over 18 hours of data collection.

VI. EXPERIMENTAL RESULTS

A. Comparative analysis of ECLAIR performance



(a) (b)

Fig. 3: Evaluation of ECLAIR and TAMER on the *Pick-and-Place* task. (a) Performance curve of the models. The expert performance (green dashed line) sets the targeted behaviours. ECLAIR quickly converges towards expert performance, while TAMER does so at a slower pace. (b) Success rates of the training session. ECLAIR is significantly more successful than TAMER.

The objective of this experiment is to evaluate the efficacy of ECLAIR in enhancing the robot's learning process from human advice. Figure 3a presents the learning progression for both ECLAIR and the baseline method, TAMER, in the pick-and-place task. ECLAIR, represented by the blue line, demonstrates rapid learning, closely mirroring the expert benchmark (dashed green line) from the initial episodes. In contrast, TAMER, depicted by the orange line, shows a gradual and slower improvement, achieving successful trials only towards the end of the training sessions. The quick convergence of ECLAIR enabled it to be over 70% more effective than TAMER across the teaching trials, $p < 0.0001$, as detailed in Figure 3b.

ECLAIR's enhanced learning is largely due to its integration of different types of human advice into the learning of the robot. TAMER, using only evaluative feedback confines the participants to only comment on the robot's past actions. This often leads to an extensive exploration by the robot, exploiting the human feedback, only when receiving a positive one upon executing a correct action by chance. ECLAIR, in contrast, integrates both corrective feedback and guidance to shape the robot's behaviour. In this setting, participants do not have to passively observe the robot's self-guided discovery of correct actions but can proactively steer it towards these optimal ones, either by correcting past mistakes or providing foresight on upcoming states. This requires less exploration from the robot, leading to a quick convergence to the optimal policy.

Figure 4, which presents a heatmap of state visitations during training, supports this interpretation. ECLAIR's heatmap displays a more condensed activity along the optimal trajectory, suggesting focused training, whereas TAMER's heatmap indicates more widespread and less focused exploration, including visits to sub-optimal states.

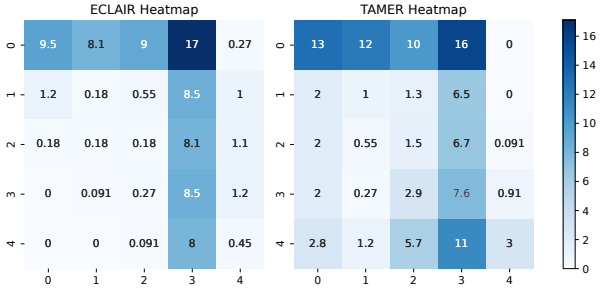


Fig. 4: Heatmap of state visitations during training. ECLAIR displays a concentrated activity along the optimal trajectory, while TAMER exhibits a broader exploration.

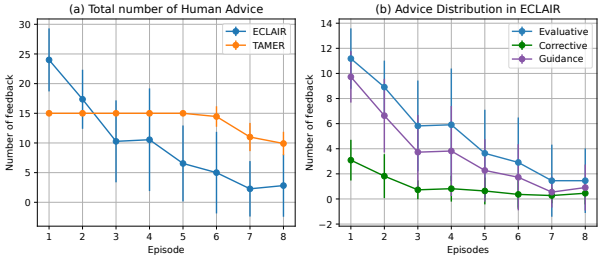


Fig. 5: Distribution of human feedback across training episodes. (a) Compares the total advice provided in the ECLAIR and TAMER models, highlighting ECLAIR's initial higher engagement that decreases over time, opposite to TAMER's consistent feedback pattern. (b) Breaks down the types of feedback used within ECLAIR, showing a preference for evaluative and guidance feedback over corrective feedback.

B. Feedback dynamics

The prior experiment validated that integrating diverse types of human advice significantly improves robot learning. This follow-up study aims to quantify the necessary amount of human inputs for achieving learning convergence and to examine how participants engage with the various feedback channels.

Figure 5a presents the feedback distribution across training episodes for both the ECLAIR and TAMER models. Notably, for ECLAIR, we aggregate all feedback types to quantify the total human advice provided during sessions. Initially, the ECLAIR model, represented by the blue line, sees significantly more input from participants compared to TAMER, denoted by the orange line. However, this trend reverses from the third episode onwards, with ECLAIR feedback converging to zero, while TAMER feedback remains consistently high until converging, where it then slightly diminishes.

These observations suggest that the volume of feedback correlates strongly with the robot's learning progress—participants tend to offer less input as the robot improves. The availability of multiple feedback channels in ECLAIR seemingly reduces the cognitive load on trainers, allowing them to offer varied and richer feedback at the initial stage of training, enabling quicker learning with less overall trainer effort.

In contrast, TAMER's single feedback channel may lead to more constant use, as it's the sole means of communication with the robot, potentially leading to its misuse for other forms of instruction, as noted in prior studies [9]. This feedback efficiency positions ECLAIR as a more effective method by minimising the need for extensive human intervention during robot training.

Additionally, we investigated participant interactions with each feedback channel within the ECLAIR model. Figure 5b depicts the categorisation of advice distributed across the teaching episodes. We observe that participants predominantly used evaluative feedback when instructing the robot, with guidance closely following, indicating a tendency among participants to employ a combination of these feedback types for teaching. The use of corrective feedback was notably less frequent. A plausible explanation for this trend is that the provision of guidance likely directed the robot towards optimal actions, which were subsequently reinforced through evaluative feedback. This sequence of interactions reduced the incidence of incorrect actions by the robot, diminishing the necessity and opportunity for participants to employ corrective feedback.

C. LLM's robustness as advice interpreter

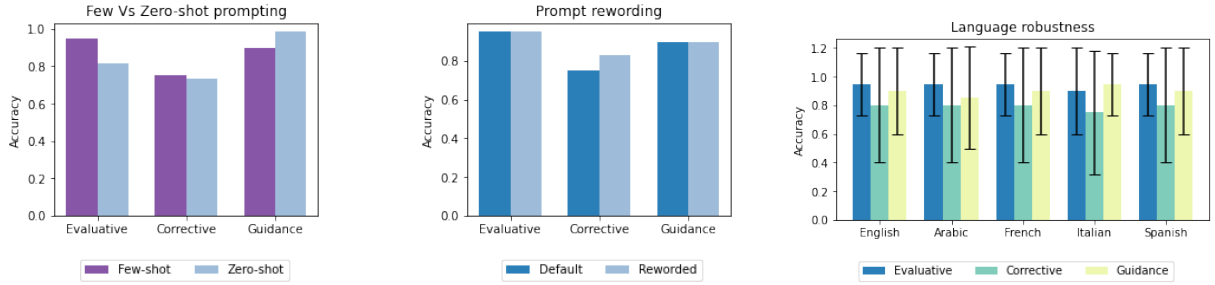
ECLAIR leverages LLMs to interpret human advice, yet LLMs performance highly depends on the input quality and prompt design [27].

In this experiment, we assess ECLAIR's adaptability to different prompting strategies, examining its performance with both few-shot and zero-shot examples. The standard prompt used in ECLAIR includes a task description, a guideline for the interpretation, and examples of desired outcomes. We compare the influence of the provided examples on interpretation accuracy. We generate a dataset of 20 instances of human advice with ground truth. We assess the label accuracy on this dataset by using the default few-shot prompt and a modified zero-shot prompt where examples were removed, for 3 seeds

Figure 6a shows that ECLAIR's performance remains stable regardless of the prompting method, suggesting LLMs' pre-training is sufficient for this task context, without requiring explicit examples.

Additionally, we investigated prompt sensitivity to wording changes following Klissarov et al.'s methodology by testing a semantically similar but reworded prompt against the standard on the same dataset, for three seeds [19]. Figure 6b illustrates the results where minimal performance differences between the prompts can be seen. This suggests that ECLAIR is robust to slight changes in prompt wording. However, it is important to note that this might be task-dependent, with more complex scenarios potentially yielding different outcomes.

Exploratory tests revealed that ECLAIR might potentially be used in multilingual interactions. To confirm this, we translate the previous advice dataset into four languages: Arabic, French, Italian, and Spanish. Figure 6c shows a



(a) (b) (c)

Fig. 6: Evaluating ECLAIR's interpretation accuracy across prompting methods and languages.

consistent label accuracy across these languages, with no significant difference. This linguistic flexibility is attributed to the LLMs' training on diverse, multilingual datasets, allowing ECLAIR to be used in different languages.

VII. CONCLUSIONS

In this study, we introduced ECLAIR, a novel framework that integrates human advice into robot learning through the interpretation of natural language feedback by Large Language Models (LLMs), subsequently applied within an RL framework. Our comprehensive evaluation demonstrated that ECLAIR significantly enhances robot learning by promoting a quick convergence to the optimal policy and requiring little human feedback through the learning, thus reducing the human teaching load. Through empirical analysis, we found ECLAIR to be robust to prompt modifications, and proficient in multiple languages. Future works should aim to expand the application of ECLAIR to more complex tasks and diverse user groups, aiming to further understand and improve robots learning from human advice.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] K. Zhu and T. Zhang, "Deep reinforcement learning based mobile robot navigation: A review," *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 674–691, 2021.

[2] J. Luo, Z. Hu, C. Xu, Y. L. Tan, J. Berg, A. Sharma, S. Schaal, C. Finn, A. Gupta, and S. Levine, "Serl: A software suite for sample-efficient robotic reinforcement learning," *arXiv preprint arXiv:2401.16013*, 2024.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[4] A. Pan, K. Bhatia, and J. Steinhardt, "The effects of reward misspecification: Mapping and mitigating misaligned models," *arXiv preprint arXiv:2201.03544*, 2022.

[5] D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, and A. Dragan, "Inverse reward design," *Advances in neural information processing systems*, vol. 30, 2017.

[6] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[7] M. Chetouani, "Interactive robot learning: an overview," *ECCAI Advanced Course on Artificial Intelligence*, pp. 140–172, 2021.

[8] A. Najar and M. Chetouani, "Reinforcement learning with human advice: a survey," *Frontiers in Robotics and AI*, vol. 8, p. 584075, 2021.

[9] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Aaai*, vol. 6. Boston, MA, 2006, pp. 1000–1005.

[10] C. Celemin and J. Ruiz-del Solar, "An interactive framework for learning continuous actions policies based on corrective feedback," *Journal of Intelligent & Robotic Systems*, vol. 95, pp. 77–97, 2019.

[11] F. Cruz, J. Twiefel, S. Magg, C. Weber, and S. Wermter, "Interactive reinforcement learning through speech guidance in a domestic scenario," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[12] J. MacGlashan, M. Littman, R. Loftin, B. Peng, D. Roberts, and M. Taylor, "Training an agent to ground commands with reward and punishment," in *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[13] J. Grizou, M. Lopes, and P.-Y. Oudeyer, "Robot learning simultaneously a task and how to interpret human instructions," in *2013 IEEE third joint international conference on development and learning and epigenetic robotics (ICDL)*. IEEE, 2013, pp. 1–8.

[14] A. Najar, O. Sigaud, and M. Chetouani, "Interactively shaping robot behaviour with unlabeled human instructions," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 2, p. 35, 2020.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[17] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[18] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," *arXiv preprint arXiv:2303.00001*, 2023.

[19] M. Klissarov, P. D'Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff, "Motif: Intrinsic motivation from artificial intelligence feedback," *arXiv preprint arXiv:2310.00166*, 2023.

[20] P. Goyal, S. Niekum, and R. J. Mooney, "Using natural language for reward shaping in reinforcement learning," *arXiv preprint arXiv:1903.02020*, 2019.

[21] T. Carta, P.-Y. Oudeyer, O. Sigaud, and S. Lamprier, "Eager: Asking and answering questions for automatic reward shaping in language-guided rl," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 478–12 490, 2022.

[22] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas, "Guiding pretraining in reinforcement learning with large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8657–8677.

[23] M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, "Teaching with rewards and punishments: Reinforcement or communication?" in *CogSci*, vol. 3, no. 3.7, 2015, p. 3.

[24] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and mdp reward." in *AAMAS*, vol. 1004. Valencia, 2012, pp. 475–482.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[26] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*, 2009, pp. 9–16.

[27] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," *arXiv preprint arXiv:2104.08786*, 2021.