Mutual Enhancement of Large Language and Reinforcement Learning Models through Bi-Directional Feedback Mechanisms: A Planning Case Study

Shangding Gu
UC Berkeley
Department of Electrical Engineering and Computer Sciences
Berkeley, CA 94720, USA
shangding.gu@berkeley.edu

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities for reinforcement learning (RL) models, such as planning and reasoning capabilities. However, the problems of LLMs and RL model collaboration still need to be solved. In this study, we employ a teacher-student learning framework to tackle these problems, specifically by offering feedback for LLMs using RL models and providing high-level information for RL models with LLMs in a cooperative multi-agent setting. Within this framework, the LLM acts as a teacher, while the RL model acts as a student. The two agents cooperatively assist each other through a process of *recursive help*, such as "I help you help I help." The LLM agent supplies abstract information to the RL agent, enabling efficient exploration and policy improvement. In turn, the RL agent offers feedback to the LLM agent, providing valuable, real-time information that helps generate more useful tokens. This bi-directional feedback loop promotes optimization, exploration, and mutual improvement for both agents, enabling them to accomplish increasingly challenging tasks. Remarkably, we propose a practical algorithm to address the problem and conduct empirical experiments to evaluate the effectiveness of our method.

1 Introduction

Large Language Models (LLMs) (OpenAI 2023; Chang et al. 2023) have shown exceptional performance across various domains. Notably, LLMs are useful for applications like robot planning (Singh et al. 2023), machine translation (Zhang, Haddow, and Birch 2023) and medicine (Thirunavukarasu et al. 2023). In parallel, RL has demonstrated remarkable capabilities in various domains, including achieving human-level performance in games such as the game of Go (Silver et al. 2016) and multi-player poker (Brown and Sandholm 2019). LLMs have been increasingly incorporated to enhance the performance of RL (Du et al. 2023; Szot et al. 2023). Likewise, RL has also been employed to augment the capabilities of LLMs, furthering their effectiveness (Ouyang et al. 2022; Gu, Knoll, and Jin 2024). Nevertheless, the effective harnessing of LLMs' latent potential in solving complex tasks, through the synergistic integration with powerful RL frameworks (Sutton and Barto 2018), remains a formidable challenge. Therefore, a critical question that emerges in this field is: How does an RL model cooperate with LLMs to perform a given task effectively?

Facilitating cooperation between RL models and LLMs requires mutually beneficial actions, leading to enhancing each model's performance. However, it is challenging to meet these needs due to RL models' and LLMs' distinct decision-making characteristics. Specifically, the capabilities of LLMs generally exceed those of RL models, indicating the importance of developing methodologies for instructing RL models to acquire high-level knowledge and ensuring RL models can deliver real-time feedback to LLMs.

In this study, to address the above challenge, we propose a teacher-student learning framework in a cooperative game, where the integration of RL models (students) and LLMs (teachers) with bi-directional feedback (Gu et al. 2023) may be an effective solution. The two models cooperatively carry out complex tasks, which can be considered a win-win collaboration, where the RL model and the LLM act as two agents, cooperating to complement, assist, and provide feedback to each other, ultimately solving the problem together.

2 Related Work

The relation between RL models and language representation is investigated in several methods (Chen et al. 2023; Gu, Knoll, and Jin 2024; Yang et al. 2021; Ouyang et al. 2022; Jiang et al. 2019; Uc-Cetina et al. 2023; Shinn et al. 2023; Zhao et al. 2023; Akyürek et al. 2023). For instance, in the work of Chen et al. (2023), they deploy LLMs as an optimization objective in an RL decision-making loop. Then, they further make LLMs reflect their decision results to refine LLMs' output using prompt engineering tips. In the work of (Yang et al. 2021), they leverage text as safety constraints for RL safe exploration. In the work of (Ouyang et al. 2022), they train LLMs to align with human values in an RL process, where the reward model is a supervised label to teach LLMs to follow human instructions.

The research most related to our study includes the works of Carta et al. (2023) and Tran and Le (2023). In the work of Carta et al. (2023), they employ LLMs as RL policies to acquire task-solving capabilities while learning new knowledge through interactive experiences. Their experimental

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

findings suggest that their method outperforms baseline approaches. However, a potential limitation of their work is the absence of instruction feedback from RL models, which may impact the overall effectiveness of their method. In the work of Tran and Le (2023), they deploy RL to train a conversational agent using a simulator and an initial text generated by a generative chat agent. Subsequently, they input the data from the RL-trained agent to the generative chat agent. Although their experiment results demonstrate that their method performs better than baselines, a concern of this approach may be the potential time consumption associated with RL training for multi-turn conversations, as each conversation may necessitate RL training requests. Additionally, achieving self-online learning for task execution could be challenging in this framework.

3 Method

In this section, we introduce a teacher-student learning framework with bi-directional feedback, wherein a synergistic partnership between an LLM and an RL model is employed to tackle tasks collaboratively. As illustrated in Figure 1, these two models operate in tandem, with mutual support, ultimately enabling successful task completion [1].

LLMs (teachers) help RL models (students): LLMs often struggle to generate instructions that fully capture detailed and precise environmental information. However, they can still provide approximate guidance to RL models, aiding the exploration process. By narrowing the exploration space and accelerating policy discovery, such guidance improves the efficiency of RL training. This highlights the potential of LLMs to mitigate challenges arising from imperfect instructions, thereby enhancing RL performance in policy optimization.

RL models (students) help LLMs (teachers): During policy execution in the RL framework, RL models benefit from the guidance provided by LLMs. In this collaborative process, RL models not only utilize but also evaluate the outputs generated by LLMs. This reciprocal interaction enables RL models to provide constructive feedback, facilitating the iterative refinement of LLM performance. As iterations progress, LLMs can better understand the environment, allowing them to generate increasingly effective guidance. This, in turn, enhances the ability of both RL models and LLMs to tackle complex tasks with greater efficiency. The iterative relationship between RL models and LLMs highlights their potential for continuous improvement and optimization. The corresponding algorithm is presented in Algorithm 1.

4 Experiments

Our experimental study is conducted within the BabyAI benchmark (Chevalier-Boisvert et al. 2019), utilizing the Lamorel framework (Carta et al. 2023) to support our investigation. Within this framework, we integrate RL instruction feedback into LLMs, establishing a feedback loop to enhance LLM performance. Specifically, we focus on the

[1]We use a TD error as an estimator for the case's advantage function.

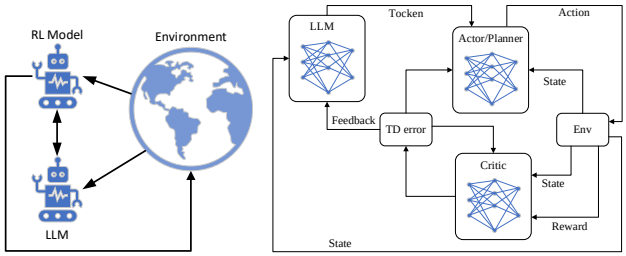


Figure 1: An LLM and an RL model collaboratively engage with an environment to accomplish complex tasks, facilitating bi-directional feedback throughout the process.

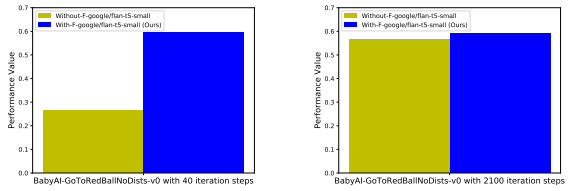
Algorithm 1: Onling Learning to make decisions with RL and LLMs.

Initial Q value Q_{θ_0}, V value V_{ϕ_0}, and advantage function value $A_0^P = Q_{\theta_0} - V_{\phi_0}$, state s_0, action a_0, token x_0.
for $t = 0, 1, \ldots, T$ **do**
  Conduct tasks with an RL model $P_{\theta_t}(a_t \mid x_t, s_t)$ and an LLM $M(x_t \mid s_t)$.
  LLM $M(x_t \mid s_t)$ provides decision information to RL model $P_{\theta_t}(a_t \mid x_t, s_t)$ by leveraging comsense capabilities and environment information.
  Estimate new RL $A_t^{P\prime}$ based on new Q value Q_{θ_t} and V value V_{ϕ_t}.
  if $A^P > A_t^{P\prime}$ **then**
    Provide negative instruction feedback to LLM, the token is worse than the last one.
  else
    Provide positive instruction feedback to LLM, the token is better than the last one.
  end if
  $A^P = A_t^{P\prime}$.
end for

GoToRedBallNoDists-v0 task, conducting experiments under two conditions: one with 40 iteration steps and another with 2100 iteration steps. We perform a comparative analysis between our proposed method and the baseline for evaluation. Our approach incorporates bidirectional interaction, where RL models provide feedback to LLMs, and LLMs supply information to RL models. In contrast, the state-of-the-art baseline, represented by the original Lamorel method, lacks this feedback mechanism from RL models to LLMs.

It is noteworthy that for our experiments, we employ the "google/flan-t5-small" model (Chung et al. 2022) as the LLM, characterized by a parameter count of 80 million. The experimental results are presented in Figure 2. These findings clearly illustrate the superior performance of our method, as quantified by the performance value metric (where higher values indicate better performance). Furthermore, our method demonstrates notably expedited convergence (one-shot/few-shot learning) when compared to the Lamorel baseline. This empirical evidence highlights the effectiveness of our approach in harnessing bi-directional

feedback between RL models and LLMs for improving performance in the context of the BabyAI benchmark.



(a) (b)

Figure 2: Experiments on BabyAI tasks (Chevalier-Boisvert et al. 2019) with 40 (a) and 2100 (b) iteration steps.

5 Conclusion

In this study, we developed a teacher-student learning framework for unlocking LLMs' powerful capabilities by leveraging an RL model with bi-directional feedback mechanisms in a cooperative game setting. To empirically assess the effectiveness of our method, we conducted experiments using the BabyAI benchmark as an assessment platform. The results of these experiments demonstrate the superior performance of our approach in comparison to the state-of-the-art baseline, highlighting its potential for substantially enhancing learning outcomes. Notably, our approach holds promise for fostering safe and robust learning systems (Gu et al. 2022), particularly in environments characterized by imperfect information. Furthermore, we hope that our findings inspire novel research directions at the intersection of LLMs and RL. As part of future work, we plan to extend our method to more challenging tasks and assess its effectiveness in complex real-world applications.

References

Akyürek, A. F.; Akyürek, E.; Madaan, A.; Kalyan, A.; Clark, P.; Wijaya, D.; and Tandon, N. 2023. RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs. *arXiv preprint arXiv:2305.08844*.

Brown, N.; and Sandholm, T. 2019. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890.

Carta, T.; Romac, C.; Wolf, T.; Lamprier, S.; Sigaud, O.; and Oudeyer, P.-Y. 2023. Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. In *ICML*, volume 202, 3676–3713. PMLR.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Zhu, K.; Chen, H.; Yang, L.; Yi, X.; Wang, C.; Wang, Y.; et al. 2023. A survey on evaluation of large language models. *arXiv:2307.03109*.

Chen, L.; Wang, L.; Dong, H.; Du, Y.; Yan, J.; Yang, F.; Li, S.; Zhao, P.; Qin, S.; Rajmohan, S.; et al. 2023. Introspective Tips: Large Language Model for In-Context Decision Making. *arXiv preprint arXiv:2305.11598*.

Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2019. BabyAI: First steps towards grounded language learning with a human in the loop. *ICLR*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.

Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; and Andreas, J. 2023. Guiding Pretraining in Reinforcement Learning with Large Language Models. In *ICML*, volume 202, 8657–8677. PMLR.

Gu, S.; Knoll, A.; and Jin, M. 2024. TeaMs-RL: Teaching LLMs to Generate Better Instruction Datasets via Reinforcement Learning. *Transactions on Machine Learning Research*.

Gu, S.; Kshirsagar, A.; Du, Y.; Chen, G.; Peters, J.; and Knoll, A. 2023. A human-centered safe robot reinforcement learning framework with interactive behaviors. *Frontiers in Neurorobotics*, 17.

Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; and Knoll, A. 2022. A review of safe reinforcement learning: Methods, theory and applications. *arXiv:2205.10330*.

Jiang, Y.; Gu, S. S.; Murphy, K. P.; and Finn, C. 2019. Language as an abstraction for hierarchical deep reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K. R.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484–489.

Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; and Garg, A. 2023. Progprompt: Generating situated robot task plans using large language models. In *ICRA*, 11523–11530. IEEE.

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

Szot, A.; Schwarzer, M.; Agrawal, H.; Mazoure, B.; Talbott, W.; Metcalf, K.; Mackraz, N.; Hjelm, D.; and Toshev, A. 2023. Large Language Models as Generalizable Policies for Embodied Tasks. *arXiv:2310.17722*.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Tran, Q.-D. L.; and Le, A.-C. 2023. Exploring bi-directional context for improved chatbot response generation using deep reinforcement learning. *Applied Sciences*, 13(8): 5041.

Uc-Cetina, V.; Navarro-Guerrero, N.; Martin-Gonzalez, A.; Weber, C.; and Wermter, S. 2023. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2): 1543–1575.

Yang, T.-Y.; Hu, M. Y.; Chow, Y.; Ramadge, P. J.; and Narasimhan, K. 2021. Safe reinforcement learning with natural language constraints. *Advances in Neural Information Processing Systems*, 34: 13794–13808.

Zhang, B.; Haddow, B.; and Birch, A. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *ICML*, volume 202, 41092–41110. PMLR.

Zhao, X.; Wang, T.; Osborn, S.; and Rios, A. 2023. BabyStories: Can Reinforcement Learning Teach Baby Language Models to Write Better Stories? *arXiv preprint arXiv:2310.16681*.