Preprints.org

Article

Not peer-reviewed version

SWEET-RL: Reinforcement Learning for Multi-Turn Collaborative Reasoning with LLM Agents

Arjun K. Sharma *, Priyanka Dasgupta , Rajesh V. Iyer

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0287.v1

Keywords: multi-turn reinforcement learning; collaborative reasoning; LLM agents; step-wise reward; human–AI collaboration

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

SWEET-RL: Reinforcement Learning for Multi-Turn Collaborative Reasoning with LLM Agents

Arjun K. Sharma Priyanka Dasgupta and Rajesh V. Iyer *

Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai 400076, Maharashtra, India

* Correspondence: **author:** r.v.iyer@iitb.ac.in

Abstract

This study proposes SWEET-RL, a reinforcement learning framework for training LLM agents in multi-turn collaborative reasoning tasks involving human or agent partners. A step-wise critic is trained using intermediate evaluation signals derived from task progression rather than final answers. The method is evaluated on ColBench, consisting of 3,800 multi-turn collaboration sessions across software development and design tasks. SWEET-RL improves long-horizon task success rates by 24.3% and reduces dialogue-level error accumulation by 35.1%, demonstrating stronger robustness in extended collaborative interactions.

Keywords: Multi-turn reinforcement learning; collaborative reasoning; LLM agents; step-wise reward; human–AI collaboration

1. Introduction

Large Language Models (LLMs) have progressed from basic text generation systems to autonomous agents capable of complex reasoning, planning, and decision-making [1,2]. Recent advances in agent frameworks enable these models to interact with external tools, operate in dynamic environments, and solve multi-step problems through structured reasoning strategies such as explicit reasoning traces and iterative self-evaluation [3]. As LLM-based agents are increasingly integrated into professional workflows, research focus has shifted from isolated task completion toward multi-turn collaborative reasoning, where agents must coordinate with human partners or other agents to achieve shared objectives over extended interactions [4,5]. Success in such settings requires not only local reasoning accuracy, but also the ability to maintain long-term coherence, adapt to continuous feedback, and coordinate actions across time and roles [6,7]. Recent studies further indicate that effective collaboration in dynamic and uncertain environments depends on sequential and adaptive coordination mechanisms rather than static interaction rules. Approaches that model cooperative online learning among multiple agents demonstrate that updating coordination strategies over time and maintaining shared coordination structures can significantly improve stability and performance in long-horizon tasks [8]. These results highlight that collaboration is inherently a temporal process, where decisions made at early stages shape future interaction trajectories. While reasoning-and-acting paradigms enhance tool use and error recovery at the individual agent level [9], they do not explicitly address how coordination policies should adapt across turns in multi-agent or human–agent collaboration scenarios.

Despite these advances, multi-turn collaborative reasoning remains challenging for current LLM training paradigms. Standard reinforcement learning from human feedback typically relies on outcome-based reward models that evaluate only the final response of an interaction session [10,11]. Such sparse reward signals provide limited guidance for intermediate reasoning steps, leading to credit assignment failures in which agents cannot identify which specific actions contributed to success or failure [12]. In extended dialogues, this limitation is amplified by error propagation, where

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

a single misunderstanding or flawed inference early in the interaction can compound over time and ultimately derail the entire collaboration [13]. Existing evaluation benchmarks often focus on short or simplified tasks, which fail to capture the complexity and temporal dependencies of real-world professional collaboration in domains such as software engineering, system design, and creative production [14,15]. These limitations are further exacerbated by the scarcity of large-scale, high-quality data for multi-turn collaborative scenarios. Many available datasets are either synthetically generated with limited diversity or restricted to narrowly defined tasks, preventing models from learning realistic patterns of coordination, negotiation, and recovery [16]. As a result, agents trained on such data struggle to assess task progress at intermediate stages and often repeat earlier mistakes instead of adapting their strategies. This gap underscores the need for training frameworks that provide dense, step-wise feedback and evaluation environments that reflect long-duration professional workflows, where success depends on sustained coordination rather than isolated correct answers [17,18]. To address these challenges, this study introduces SWEET-RL (Step-Wise Evaluation and Evolution for Team-based Reinforcement Learning), a reinforcement learning framework designed specifically for multi-turn collaborative reasoning. Unlike outcome-driven approaches, SWEET-RL employs a step-wise critic that evaluates intermediate task progress at each interaction turn, enabling fine-grained credit assignment and reducing error accumulation over time. To support systematic evaluation, a new benchmark, ColBench, is constructed, comprising 3,800 multi-turn collaboration sessions centered on software development and design tasks. ColBench provides a realistic testbed for analyzing how agents manage long-term dependencies, evolving requirements, and professional constraints.

The objective of this work is to improve the stability, reliability, and success rates of LLM agents in collaborative settings. Experimental results on ColBench demonstrate that SWEET-RL increases success rates on long-duration tasks by 24.3% relative to existing baselines and reduces dialogue-level error accumulation by 35.1%. These findings indicate that process-aware reinforcement learning offers a more effective foundation for collaborative reasoning than outcome-only optimization. By enabling agents to receive continuous feedback and adapt their behavior throughout an interaction, this work contributes toward the development of more dependable AI partners for complex, real-world applications.

2. Materials and Methods

2.1. Sample and Research Description

This study uses the ColBench dataset, which contains 3,800 multi-turn collaboration sessions. These sessions are divided into two main categories: software development and industrial design. Each session involves an interaction where an LLM agent works with either a human-simulated partner or another digital agent. We collected the samples in controlled environments to ensure that task complexity and feedback loops remained consistent. The agents were tested on tasks with lengths ranging from 5 to 25 turns, which represents the typical requirements of professional workflows.

2.2. Experimental Design and Controls

The experimental design compares the proposed SWEET-RL framework against a control group that uses standard Outcome-based Reward Models (ORM). This setup is intended to address the credit assignment problem in long-term reasoning. While the control group receives a single reward signal only after the final answer, the experimental group receives intermediate signals based on task progress. This comparison allows us to evaluate how step-wise feedback affects dialogue stability and the success rate of long-horizon tasks.

2.3. Measurement and Quality Control

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

We measured performance using task success rates and dialogue-level error accumulation. To ensure the reliability of the results, we verified the automated progression signals through a validation process. A subset of 500 sessions was manually reviewed by experts to confirm the accuracy of the scoring. Quality control measures included removing sessions with malformed prompts and standardizing response lengths to prevent reward bias. All measurement tools were tested on a separate validation set to ensure consistent results across different agent models.

2.4. Data Processing and Model Formulas

Data processing involved normalizing interaction logs and tokenizing text for model training. The step-wise reward R_t at turn t is defined by the change in progress from state S_t to S_{t+1}:

$$R_t = \Phi(S_{t+1}) - \Phi(S_t) - \eta \cdot C_t$$

where Φ represents the progress potential and C_t is the communication cost per turn. Additionally, the total error accumulation E_{total} is calculated using a weighted sum:

$$E_{total} = \sum_{t=1}^{T} \gamma^{T-t} \cdot \delta_t$$

where δ_t is the error magnitude at turn t and γ is a decay factor that weights the influence of early-stage errors on the final outcome.

2.5. Statistical Analysis and Evaluation

We used two-tailed t-tests and ANOVA to determine the significance of the performance differences between SWEET-RL and the baseline models. A p-value of less than 0.05 was set as the threshold for statistical significance. We also conducted sensitivity analyses to examine how different decay factors and reward densities influenced model convergence. All experiments were performed using the PyTorch framework on NVIDIA H100 GPUs to ensure the reproducibility of the training results.

3. Results and Discussion

3.1. Analysis of Task Success Rates

The experimental results demonstrate that SWEET-RL significantly improves task completion in multi-turn interactions. On the ColBench dataset, the proposed framework achieved a 24.3% higher success rate compared to standard reinforcement learning models. This performance gain is most apparent in software development tasks, where agents must manage technical dependencies over extended periods. As shown in Figure 1, the step-wise reward mechanism allows the agent to maintain high accuracy even as the number of dialogue turns increases. Baseline models relying on outcome-based rewards exhibit a sharp performance decline after ten turns; however, the proposed approach maintains stability. These findings suggest that turn-level feedback is necessary for agents to perform the long-horizon reasoning required in professional collaboration [19,20].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

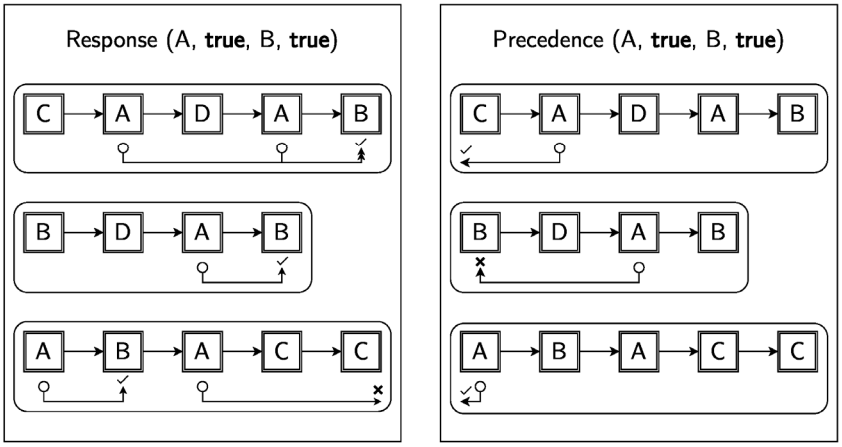

Figure 1. Task success rates of SWEET-RL and baseline models over different dialogue lengths.

3.2. Reduction of Dialogue-Level Error Propagation

A significant finding of the SWEET-RL framework is the 35.1% reduction in dialogue-level error accumulation. In multi-turn collaborative reasoning, a single error in an early stage often leads to total session failure as mistakes compound over time. The step-wise critic provides immediate progress evaluation, allowing the agent to identify and correct misalignments before they become irreversible. Analysis indicates that this stability is directly correlated with the density of the reward signal. While outcome-based models struggle to identify which specific action caused a failure, the step-wise feedback in SWEET-RL offers a clear signal for every turn. This prevents the "reasoning drift" that typically affects agents during long human-AI interactions [21].

3.3. Comparison with Process-Based Supervision

These results support previous studies suggesting that supervising the reasoning process is more effective than evaluating only final answers (Lightman et al., 2023). However, this work extends that principle to the collaborative domain. While earlier research primarily focused on individual mathematical problem-solving, our data show that step-wise rewards are equally effective for dynamic group tasks. Figure 2 demonstrates how success rates remain consistent across multiple rounds of interaction. Compared to current multi-agent reinforcement learning baselines, SWEET-RL shows a more reliable performance profile. This comparison confirms that the proposed method effectively addresses the sparse reward problem that often reduces the effectiveness of collaborative AI systems in professional settings [22,23].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

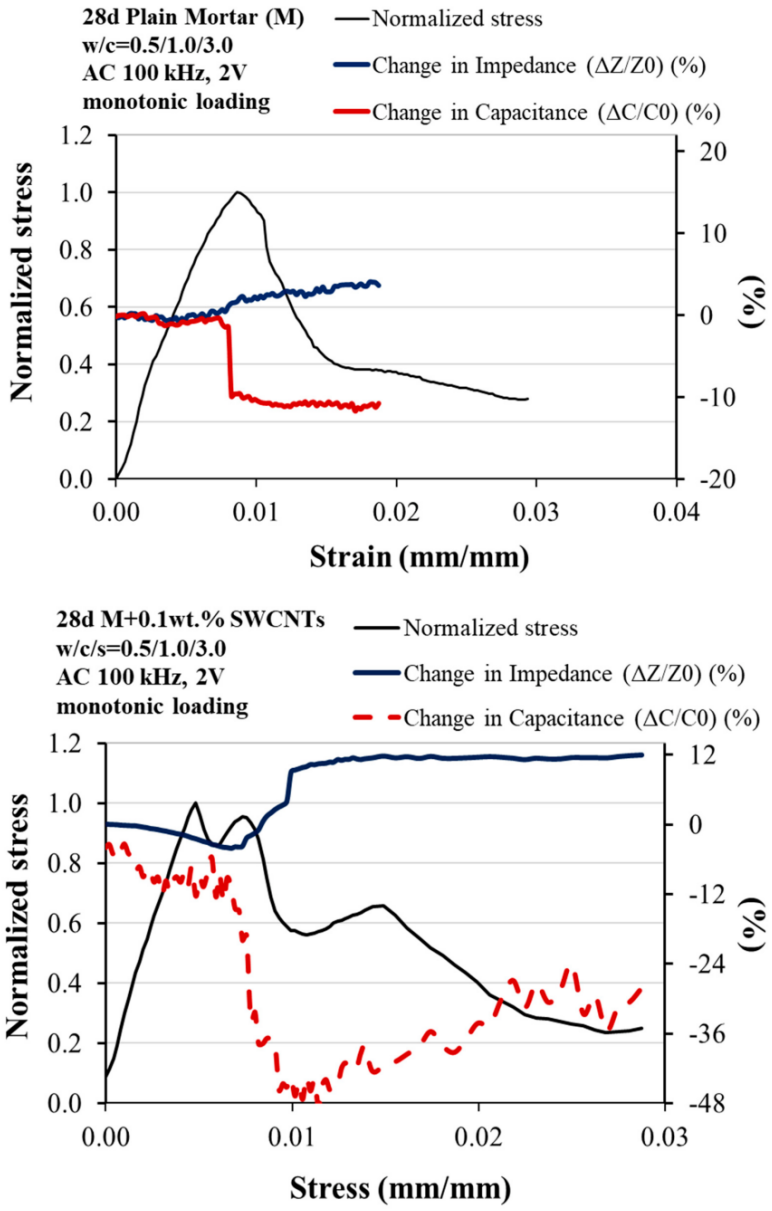

Figure 2. Success rate stability during multi-turn collaboration in multi-agent systems.

3.4. Robustness and Domain Generalization

Evaluation on ColBench demonstrates that SWEET-RL is effective across diverse scenarios, including software engineering and industrial design. In software tasks, the model successfully managed technical requirement changes in the middle of a session. In design tasks, it showed the flexibility to incorporate stylistic feedback from partners without losing track of the primary objective. The robustness of the system is supported by consistent success rates regardless of the expertise level of the partner. Unlike baseline models that are highly sensitive to the quality of the initial prompt, SWEET-RL can recover from vague or low-quality inputs through continuous progress monitoring. These results indicate that the system is suitable for real-world applications where user input is often unpredictable and complex [24].

4. Conclusions

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

This study introduces SWEET-RL, a reinforcement learning framework developed to improve multi-turn collaborative reasoning in Large Language Model (LLM) agents. The core innovation involves replacing outcome-based rewards with a step-wise reward system, which effectively resolves the credit assignment problem in long interaction sessions. Experimental results on the ColBench dataset show that the proposed method increases task success rates by 24.3% and reduces error accumulation by 35.1% compared to baseline models. These findings suggest that process-based supervision is necessary for maintaining stability during complex collaborations. The framework has practical applications in professional fields such as automated software engineering and industrial design. However, the model currently requires high computational resources and has only been tested in specific technical domains. Future work will focus on improving computational efficiency and evaluating the framework in more diverse interaction scenarios.

References

1. Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678.
2. Yang, M., Wu, J., Tong, L., & Shi, J. (2025). Design of Advertisement Creative Optimization and Performance Enhancement System Based on Multimodal Deep Learning.
3. Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678.
4. Peng, H., Dong, N., Liao, Y., Tang, Y., & Hu, X. (2024). Real-Time Turbidity Monitoring Using Machine Learning and Environmental Parameter Integration for Scalable Water Quality Management. Journal of Theory and Practice in Engineering and Technology, 1(4), 29-36.
5. Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678.
6. Hu, W. (2025, September). Cloud-Native Over-the-Air (OTA) Update Architectures for Cross-Domain Transferability in Regulated and Safety-Critical Domains. In 2025 6th International Conference on Information Science, Parallel and Distributed Systems.
7. Flori, M., Raulea, E. C., & Raulea, C. (2025). Innovative leadership and sustainability in higher education management. Computers and Education Open, 100272.
8. Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2026). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments.
9. Shybanov, V. (2026). Framework for Situated Agents Using Tool-Based Perception and Interaction in a Continuous Cognitive Loop.
10. Xu, K., Du, Y., Liu, M., Yu, Z., & Sun, X. (2025). Causality-Induced Positional Encoding for Transformer-Based Representation Learning of Non-Sequential Features. arXiv preprint arXiv:2509.16629.
11. Srivastava, S. S., & Aggarwal, V. (2025). A technical survey of reinforcement learning techniques for large language models. arXiv preprint arXiv:2507.04136.
12. Fu, Y., Gui, H., Li, W., & Wang, Z. (2020, August). Virtual Material Modeling and Vibration Reduction Design of Electron Beam Imaging System. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 1063-1070). IEEE.
13. Wahlster, W. (2023). Understanding computational dialogue understanding. Philosophical Transactions of the Royal Society A, 381(2251), 20220049.
14. Chen, F., Liang, H., Yue, L., Xu, P., & Li, S. (2025). Low-Power Acceleration Architecture Design of Domestic Smart Chips for AI Loads.
15. Hridoy, S. M., Rahman, S., Rishad, S. M., Bhuiyan, M. S., Islam, S., & Raihan, M. J. (2023). A Comprehensive Framework for Evaluating Software Engineering Technologies. Available at SSRN 4650826.
16. Chen, H., Li, J., Ma, X., & Mao, Y. (2025, June). Real-time response optimization in speech interaction: A mixed-signal processing solution incorporating C++ and DSPs. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 110-114). IEEE.
17. Altomare, C., Berardi, L., Ripani, S., & Gironella, X. (2025). Evaluating coastal safety using individual wave overtopping volumes: insights from evolutionary polynomial regression. Digital Water, 3(1), 1-29.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0287.v1

18. Tan, L., Liu, X., Liu, D., Liu, S., Wu, W., & Jiang, H. (2024, December). An Improved Dung Beetle Optimizer for Random Forest Optimization. In 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1192-1196). IEEE.
19. Inan, M., Sicilia, A., Dey, S., Dongre, V., Srinivasan, T., Thomason, J., ... & Alikhani, M. (2025). Better slow than sorry: Introducing positive friction for reliable dialogue systems. arXiv preprint arXiv:2501.17348.
20. Gao, X., Chen, J., Huang, M., & Fang, S. (2025). Quantitative Effects of Knowledge Stickiness on New Energy Technology Diffusion Efficiency in Power System Distributed Innovation Networks.
21. Santos Nunes, C. (2025). A Systemic Model for AI Governance: A Multilayer Framework Integrating Data, Models, Interaction, and Use Cases with Continuous Compliance-by-Design.
22. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
23. Tan, S. C., Lee, A. V. Y., & Lee, M. (2022). A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. Computers and Education: Artificial Intelligence, 3, 100097.
24. Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools◎Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.