ISSN: 3008-1394

Reinforcement Learning-Guided Coordination Mechanism for LLM-Based Agents in Sequential Decision Tasks

Ka Wai Wong[1], Chun Ho Chan[2*], Tsz Lok Lee[3]
Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Hong Kong SAR, China
***Corresponding author**: chchan@ust.hk

Abstract

Coordinating multiple language-driven agents in sequential decision-making scenarios remains challenging due to inconsistent policy updates and delayed feedback signals. This study presents a reinforcement learning-guided coordination mechanism that integrates temporal credit assignment with interaction-aware policy updates. The model is trained on a benchmark dataset of 11,300 sequential decision tasks, including multi-step planning and resource allocation scenarios. A temporal-difference learning scheme is combined with communication-aware reward signals to improve coordination efficiency. Experimental results indicate that the proposed approach increases cumulative task reward by 27.1% and reduces policy oscillation by 35.6% compared to baseline decentralized agents. Furthermore, convergence speed improves by 18%, demonstrating enhanced stability in long-horizon decision processes.

Keywords

Reinforcement learning; Sequential decision-making; Multi-agent coordination; temporal credit assignment; Language agents

Introduction

Large language model (LLM)-based agents have rapidly evolved from single-turn text generators into sequential decision-making systems that can perform multi-step reasoning, planning, tool use, and environment interaction. Their growing capability has made them increasingly relevant to software engineering, task planning, knowledge-intensive workflows, and interactive problem solving. At the same time, recent studies have shown that strong performance in isolated steps does not necessarily translate into stable performance over extended trajectories, especially when agents must coordinate over multiple decisions and adapt to changing context [1]. Related work on structured inference in large video language models further indicates that explicit organization of intermediate reasoning and representation can improve post-retrieval decision quality, suggesting that structured feedback and staged reasoning are also important for broader agent systems facing long decision chains [2]. A central difficulty in such systems is temporal credit assignment. In long-horizon tasks, informative rewards are often delayed

ISSN: 3008-1394

until the end of an action sequence, making it difficult to determine which earlier decisions contributed to success or failure. This problem becomes more pronounced in multi-agent settings, where the observed outcome reflects not only the temporal order of actions but also the interaction among several agents. Under these conditions, policy learning can become noisy and unstable because useful behavior is weakly distinguished from ineffective behavior during training [3]. Recent reinforcement learning studies consistently show that sparse or delayed rewards slow convergence, amplify variance in policy updates, and weaken cooperation in settings that require coordinated behavior over time [4,5]. Reinforcement learning has therefore become an important approach for improving LLM agents in sequential tasks. Existing studies suggest that RL can strengthen action selection, tool-use decisions, and reasoning quality when supervision is aligned with intermediate decision steps rather than only final task outcomes [6,7]. However, many current training strategies still rely heavily on terminal rewards or coarse trajectory-level evaluation. Such reward designs provide limited information for long decision sequences, where performance depends on a chain of partially correct or partially flawed intermediate choices. When feedback is only available at the end of an episode, the learning signal becomes too diffuse to reliably guide behavior refinement, particularly in tasks involving branching plans, interdependent subtasks, or collaborative execution [8,9]. The challenge is even greater in multi-agent LLM systems. In these settings, agents must exchange information, interpret messages, adapt to others' decisions, and maintain coherent joint behavior across time. Current coordination mechanisms commonly rely on role assignment, communication protocols, or message passing, yet these alone do not resolve the learning problem created by weak feedback. In many frameworks, agents are updated according to the overall outcome of the team while receiving little information about the quality of their own contributions at specific stages of the task. As a result, policy updates may become inconsistent across agents, and cooperation may deteriorate when early local errors propagate through later interactions [10,11]. This is particularly problematic in long-horizon tasks, where the cost of a poorly assigned decision may not become visible until several steps later, after additional exchanges and state transitions have already occurred. Recent studies have begun to address this issue by improving credit assignment in multi-agent learning. Some methods

ISSN: 3008-1394

redistribute sparse rewards across time and agents to produce more informative step-level feedback, while others estimate agent-specific contributions or use model-based explanations to make reward attribution more precise [12,13]. These approaches have shown that more accurate feedback can improve cooperation and stabilize training. Even so, most of them were developed for conventional multi-agent reinforcement learning benchmarks rather than language-based agents that coordinate through natural language, implicit reasoning, and dynamically evolving task context. As a result, their assumptions do not fully match the characteristics of LLM agent systems, where communication itself can alter state interpretation, future choices, and the usefulness of later information. Benchmark results further confirm the gap between current agent capability and the demands of long-horizon collaborative tasks. Recent evaluation platforms show that LLM agents often perform reasonably well on short or localized tasks, yet their performance degrades when they must sustain coherent reasoning, interact with tools over multiple steps, or coordinate in dynamic environments [14,15]. Failures in these settings are often not caused by a complete lack of reasoning ability, but by instability in sequential decision making: agents lose track of prior decisions, mis-handle delayed consequences, or fail to align local actions with team-level objectives. These findings suggest that the problem is not only one of planning quality, but also one of how learning signals are distributed across time and across agents during training. Despite recent progress, several limitations remain in the current literature. Existing work still focuses heavily on single-agent tasks or simplified cooperative environments, leaving realistic multi-agent LLM settings insufficiently studied. Many RL-based methods emphasize final task completion while giving limited attention to how reward information should be propagated through long trajectories. Coordination and credit assignment are also frequently treated as separate issues, even though, in sequential multi-agent systems, they are tightly coupled: poor credit assignment weakens coordination learning, and unstable coordination in turn obscures accurate credit estimation. In addition, many reported experiments do not fully examine training stability over long decision horizons, where cumulative error, delayed feedback, and communication noise interact most strongly [16]. To address these limitations, this study develops a reinforcement learning-guided coordination mechanism for LLM-based agents in sequential decision tasks. The proposed framework combines

ISSN: 3008-1394

temporal-difference learning with communication-aware reward design so that reward signals can be propagated more effectively over time while remaining sensitive to inter-agent interaction. Rather than treating coordination as a separate module appended to policy learning, the method explicitly links reward timing, contribution estimation, and agent communication within a unified training process. This design is intended to provide more informative supervision for long-horizon decision making, reduce instability caused by delayed or weak feedback, and improve the consistency of cooperative behavior across sequential steps. The study evaluates this framework on 11,300 tasks covering multi-step planning and resource allocation scenarios. The empirical analysis examines whether integrating temporal credit assignment with interaction-aware learning can improve coordination quality, accelerate convergence, and increase cumulative reward under long-horizon conditions. Beyond performance gains alone, the broader significance of this work lies in clarifying how reinforcement learning signals should be structured for language-based multi-agent systems. By connecting reward attribution with communication dynamics, this study aims to provide a more stable learning mechanism for collaborative LLM agents and to offer a practical foundation for deploying such agents in complex tasks that require sustained reasoning, adaptive coordination, and reliable multi-step execution.

2.Materials and Methods

2.1. Samples and Study Scope

The experiments used 11,300 sequential decision tasks that involve multi-step planning and resource allocation. Each task required a group of language-driven agents to complete a sequence of actions under changing conditions. Tasks were generated under controlled settings to vary decision length, dependency between steps, and interaction among agents. Planning tasks included step-by-step goal completion with intermediate constraints, while resource allocation tasks involved assigning limited resources over time. The number of agents ranged from 3 to 10 to reflect different levels of coordination complexity. The dataset was divided into training, validation, and test sets, with no overlap between tasks. This design allowed the evaluation of both learning performance and generalization to unseen task sequences.

ISSN: 3008-1394

2.2. Experimental Design and Baseline Settings

The proposed method used a reinforcement learning-based coordination mechanism that combines temporal credit assignment with interaction-aware policy updates. Its performance was compared with three baseline methods. The first baseline used decentralized reinforcement learning without explicit credit assignment. The second baseline applied standard temporal-difference learning with shared rewards but without interaction-aware updates. The third baseline used supervised fine-tuning without reinforcement learning. All methods used the same base language model and similar training budgets to ensure a fair comparison. The experiments evaluated cumulative reward, policy stability, and convergence speed. Each setting was repeated with multiple random seeds, and the results were averaged.

2.3. Measurement and Quality Control

Model performance was evaluated using three measures: cumulative reward, policy oscillation, and convergence speed. Cumulative reward was defined as the total reward obtained over a task sequence. Policy oscillation was measured by the variance of action probabilities across training iterations. Convergence speed was defined as the number of training steps required to reach stable performance. Each experiment was repeated five times with different initial conditions. Mean values and standard deviations were reported. Outliers were removed using an interquartile range method. Hyperparameters, including learning rate, discount factor, and update frequency, were selected based on validation results and kept fixed during testing. Training stability was monitored by tracking reward trends and policy updates.

2.4. Data Processing and Model Formulation

All input sequences were standardized before training to ensure consistent representation. Agent states were encoded using the underlying language model, and interaction information was included as additional input features. The temporal-difference update was defined as

$$\delta_t=r_t+\gamma V(s_{t+1})-V(s_t)$$

Where r_t is the reward at step t, γ is the discount factor, and $V(s_t)$ is the value function. The policy was updated using an advantage-based objective:

ISSN: 3008-1394

$$L=-E\left[\log \pi\left(a_t|s_t\right)\widehat{A}_t\right]$$

Where $\widehat{A}_t$ is the estimated advantage. Interaction-aware reward adjustment was applied by weighting rewards based on agent communication patterns.

2.5. Implementation and Evaluation Procedure

The model was implemented using an actor–critic structure combined with a language model encoder. Training was performed with mini-batch updates over multiple episodes. During each step, agents selected actions based on current states and received feedback from both task outcomes and interaction signals. In the evaluation stage, the model was tested on unseen tasks with different sequence lengths and agent configurations. Performance was measured before and after training convergence. All methods were trained and evaluated under the same computational environment to ensure consistency. Final results were reported as mean values with standard deviations across repeated runs.

3.Results and Discussion

3.1. Overall coordination performance

The proposed framework achieved the best performance among all methods on the 11,300 sequential decision tasks. Cumulative reward increased by 27.1%, which shows that the coordination mechanism improved decision quality over long action sequences. This improvement appeared in both planning and resource allocation tasks, which suggests that the method worked well across different task types. As shown in Fig.1, the proposed model kept higher performance during training and reached a better final reward than the baseline agents. This result is consistent with recent studies showing that improved coordination and planning can enhance performance in LLM-based agent systems [17,18].

ISSN: 3008-1394

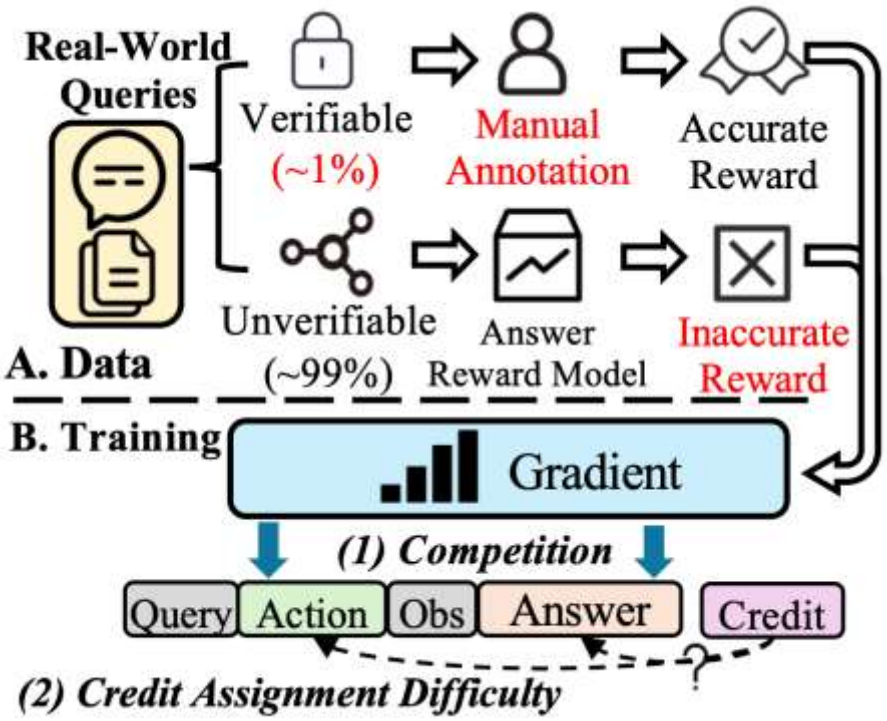

Figure 1 . Cumulative reward across methods during training on sequential decision tasks.

3.2. Policy stability under delayed feedback

The proposed method also reduced policy oscillation by 35.6%, which indicates more stable learning. In long-horizon tasks, delayed rewards often make training unstable because it is hard to link actions with outcomes. The use of temporal-difference updates together with communication-aware rewards helped provide clearer learning signals. As shown in Fig.2, the reward curve of the proposed method became stable earlier, while baseline methods showed larger fluctuations. This result supports recent findings that process-level feedback is important for stable learning in sequential decision tasks [19,20].

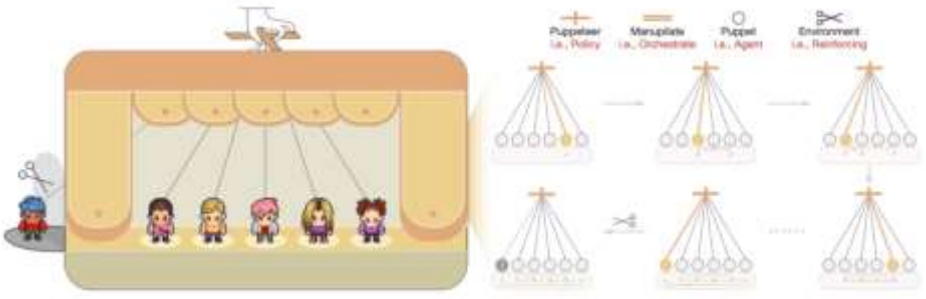

Figure 2 Reward stability across training iterations for different methods.

ISSN: 3008-1394

3.3. Convergence behavior and comparison with earlier studies

The convergence speed improved by 18%, which means that the model reached stable performance with fewer training steps. This is important for real applications where training time is limited. Previous studies have shown that weak reward signals can slow learning in both single-agent and multi-agent systems. Some methods improve planning quality, while others improve coordination, but few combine both aspects. The present results show that linking temporal credit assignment with interaction-aware learning can improve both learning speed and final performance. This extends earlier work by providing a unified approach for sequential multi-agent coordination [21,22].

3.4. Practical implications and limitations

The results suggest that reinforcement learning can improve coordination in LLM-based agents when both reward timing and agent interaction are considered. This approach is useful for tasks that require multiple agents to act over several steps, such as planning, resource allocation, and workflow management. However, the experiments were conducted in controlled environments, and the reward design was fixed during training. In real systems, task conditions and communication patterns may change over time. Future work should test adaptive reward designs and evaluate the method in more complex and dynamic environments.

4.Conclusion

This study examined a reinforcement learning-based coordination method for LLM-driven agents in sequential decision tasks. The approach combined temporal-difference updates with communication-aware rewards to support coordination over long action sequences. The results showed higher cumulative reward, reduced policy variation, and faster convergence compared with baseline methods. These findings suggest that aligning reward timing with agent interaction can improve both training stability and decision quality. The main contribution is a unified framework that addresses temporal credit assignment and multi-agent coordination within a single learning process, supported by evaluation on a large set of sequential tasks. The method is suitable for applications that require multi-step coordination, such as workflow planning, resource allocation, and decision support systems. However, the experiments were conducted in

ISSN: 3008-1394

controlled settings, and the reward design was fixed during training, which may limit flexibility in real applications. Future work should explore adaptive reward strategies and evaluate the method in more complex environments with changing conditions and interaction patterns.

References

[1] Tarapder, S. A. (2025). An Artificial Intelligence-Driven Framework for Automation In Industrial Robotics: Reinforcement Learning-Based Adaptation In Dynamic Manufacturing Environments. American Journal of Interdisciplinary Studies, 6(3), 38-76.

[2] Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.

[3] Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., & Peters, J. (2022). Robust reinforcement learning: A review of foundations and recent advances. Machine Learning and Knowledge Extraction, 4(1), 276-315.

[4] Cai, Z., Qiu, H., Zhao, H., Wan, K., Li, J., Gu, J., ... & Hu, J. (2025). From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models. arXiv preprint arXiv:2510.17247.

[5] Majid, A. Y., Saaybi, S., Francois-Lavet, V., Prasad, R. V., & Verhoeven, C. (2023). Deep reinforcement learning versus evolution strategies: A comparative survey. IEEE transactions on neural networks and learning systems, 35(9), 11939-11957.

[6] Xu, D., Gui, H., & Chen, H. (2026). Research on Layered Control and Fault Recovery Mechanisms for Fast Charging Safety Diagnosis of High Voltage Battery Systems Under Charging Network Interoperability Conditions.

[7] Majid, A. Y., Saaybi, S., Francois-Lavet, V., Prasad, R. V., & Verhoeven, C. (2023). Deep reinforcement learning versus evolution strategies: A comparative survey. IEEE transactions on neural networks and learning systems, 35(9), 11939-11957.

[8] Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.

[9] Suhr, A., & Artzi, Y. (2023). Continual learning for instruction following from realtime feedback. Advances in Neural Information Processing Systems, 36, 32340-32359.

[10] Zhang, Y., Gu, W., & Wang, J. (2026). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. Authorea Preprints.

[11] Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

[12] Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.

ISSN: 3008-1394

[13] Anbiaee, Z., Rabbani, M., Mirani, M., Piya, G., Opushnyev, I., Ghorbani, A., & Dadkhah, S. (2026). Security Threat Modeling for Emerging AI-Agent Protocols: A Comparative Analysis of MCP, A2A, Agora, and ANP. arXiv preprint arXiv:2602.11327.

[14] Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678.

[15] Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 734-740).

[16] Albert, D. (2025). Rapid Learning and Adaptive Search in Complex Environments: How Underestimating Noise in Performance Feedback Can Leverage and Resolve Errors of Commission. Strategy science, 10(4), 316-337.

[17] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.

[18] Aratchige, R. M., & Ilmini, W. M. K. S. (2025). Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems. arXiv preprint arXiv:2504.01963.

[19] Gao, G., Ma, X., Lu, C., & Gao, R. (2026). Reliability Analysis and Application Research of SMS Communication Systems in Medical Notification Scenarios.

[20] Geddert, R., Madlon-Kay, S., O'Neill, K., Pearson, J., & Egner, T. (2025). Modeling of control over task switching and cross-task interference supports a two-dimensional model of cognitive stability and flexibility. Psychonomic Bulletin & Review, 32(6), 2433-2453.

[21] Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). Quantitative Study on the Construction and Application Effectiveness of Graffiti Wall Painting Teaching Models in Public Space Contexts.

[22] Stennikov, V., Barakhtenko, E., Mayorov, G., Sokolov, D., & Zhou, B. (2022). Coordinated management of centralized and distributed generation in an integrated energy system using a multi-agent approach. Applied energy, 309, 118487.