Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 5, No. 5, 2025

Knowledge-Informed Policy Structuring for Multi-Agent Collaboration Using Language Models

Yumeng Ma[1], Guohui Cai[2], Fan Guo[3], Zhou Fang[4], Xiaokai Wang[5]
[1]Arizona State University, Tempe, USA
[2]Illinois Institute of Technology, Chicago, USA
[3]Illinois Institute of Technology, Chicago, USA
[4]Georgia Institute of Technology, Atlanta, USA
[5]Santa Clara University, Santa Clara, USA
*Corresponding Author: Xiaokai Wang; shawnxkwang@gmail.com

Abstract:

This paper proposes a multi-agent cooperative reinforcement learning method guided by large language model strategies. It addresses issues in traditional multi-agent reinforcement learning, such as policy instability and low exploration efficiency. By introducing strategy guidance generated by large language models, the method helps agents converge quickly to optimal cooperative policies. First, a fusion mechanism between the language model and the reinforcement learning framework is established. A dynamic guidance mechanism is designed to adaptively adjust the strength of guidance. This enhances policy stability and increases cooperation success rates across the system. Experiments show that the proposed method outperforms traditional joint policy approaches under various task complexities, agent scales, and guidance strength settings. These results validate the effectiveness of the proposed strategy guidance mechanism.

Keywords:

Multi-agent reinforcement learning; large language model; strategy guidance; dynamic guidance mechanism

1. Introduction

Multi-Agent Reinforcement Learning (MARL), as a crucial subfield of reinforcement learning, demonstrates significant potential in complex task modeling, agent collaboration control, and intelligent decision system construction. As real-world environments become increasingly complex, single-agent learning and decision-making become insufficient for handling high-dimensional and highly coupled tasks. As a result, multi-agent collaboration has become a key focus in artificial intelligence research. However, organizing effective collaboration strategies among agents, avoiding unstable game dynamics, and improving system convergence and task completion remain significant [1, 2]. On one hand, although agents may share the same environment, divergences in their perspectives and objectives often cause strategy conflicts and information asymmetry, leading to unstable training. On the other hand, in the absence of high-quality prior knowledge, cooperation among agents heavily relies on lengthy and inefficient exploration, hindering effective [3, 4].

In recent years, large language models (LLMs) have achieved breakthroughs in natural language understanding and generation[5]. Their abilities in contextual comprehension, structured information organization, and strategic generation have drawn attention in the reinforcement learning domain. LLMs can bridge the gap between unstructured information and decision variables through their knowledge representation capabilities. As external policy generators, they provide agents with globally informed and

knowledge-guided behavioral suggestions[6]. This language-model-driven policy guidance offers a new approach to reinforcement learning: agents can enhance their decision-making by leveraging external knowledge from language models, reducing learning costs and improving policy optimization. Especially in multi-agent settings, LLMs can act as "coordinators," enabling organized and structured collaboration in complex tasks while mitigating issues like strategy conflicts and unstable interactions[7].

Integrating LLMs into multi-agent systems extends their applications in cognitive computing and behavioral modeling. It also introduces new challenges and opportunities for traditional reinforcement learning paradigms. One challenge is how to effectively align language models with environmental states and translate natural language into executable strategies. Another is how to fuse policy suggestions from language models with locally observed strategies of individual agents to achieve robust and efficient collaboration[8]. Moreover, LLMs offer scalability and transferability, supporting the development of general-purpose collaboration systems across tasks and domains. This cross-modal collaboration approach breaks the conventional perception-centered decision architecture, building a tighter loop among cognition, language, and behavior.

This study aims to establish an efficient integration mechanism between reinforcement learning and LLMs. From the perspective of policy guidance, it explores how language models can enhance collaboration in multi-agent systems. By incorporating policy priors generated by LLMs, agents can bypass inefficient exploration and reach faster policy convergence. Meanwhile, LLMs can serve as sources of knowledge distillation, abstracting global training experiences into general strategy templates to guide decision-making in specific contexts. Furthermore, in real-world applications such as disaster response, swarm robotics, and complex strategic games, policy-guided multi-agent systems can demonstrate higher flexibility, reliability, and responsiveness, enhancing both intelligence and practical value.

In conclusion, policy-guided multi-agent reinforcement learning using LLMs represents a cutting-edge direction in multimodal fusion and collaborative learning. It holds both theoretical and practical significance. On one hand, it promotes a shift from inefficient, spontaneous exploration to efficient, knowledge-driven collaboration in MARL. On the other, it provides a new technical route for building intelligent systems with capabilities in language understanding, task collaboration, and autonomous decision-making. This study aims to uncover the policy-level guiding potential of LLMs, constructing a semantic-policy fusion learning framework to expand the boundaries of agent collaboration and lay the theoretical foundation for future deployment in general artificial intelligence systems.

2. Related work

Multi-agent reinforcement learning has made significant progress in recent years. Representative approaches include the Centralized Training with Decentralized Execution (CTDE) framework, joint policy gradient optimization, and attention-based cooperative strategy modeling. These methods introduce shared global information or enhance inter-agent cooperation through communication mechanisms [9]. They help alleviate certain issues such as non-stationarity and strategic conflicts. However, most existing approaches rely on intensive interactions or frequent synchronization. This results in poor scalability and slow convergence, especially in large-scale or heterogeneous agent systems. Moreover, individual policies typically depend on long-term interaction accumulation, lacking effective prior knowledge guidance. This limits the system's adaptability in complex environments[10].

Recently, the auxiliary role of language models in agent decision-making has gained increasing attention. Some studies have explored the use of pretrained language models to generate task prompts, environment descriptions, or behavioral templates[11]. These are used to improve the generalization of reinforcement

learning agents in complex semantic tasks. For instance, methods such as Prompt-based RL and Language-conditioned RL adopt language as an intermediate bridge to guide policy learning or action selection. These approaches show good adaptability and task transferability[12]. However, such research mainly focuses on single-agent settings. Their application and influence in multi-agent cooperative systems remain underexplored. In particular, when agents face policy coupling, local information asymmetry, and reinforcement-dependent actions, the policy guidance mechanism of language models lacks systematic modeling and empirical study[13].

In addition, most existing studies that combine language models with reinforcement learning focus on isolated components. These include policy initialization, behavior generation, or task decomposition. There is a lack of holistic design for policy guidance from a collaboration perspective. A unified framework to systematically integrate the reasoning capabilities of language models with multi-agent cooperative strategies has not yet been established. Meanwhile, effective coordination mechanisms are still missing for resolving conflicts between strategies generated by language models and behaviors learned by agents through environmental feedback. Therefore, research on policy modeling, coordination design, and convergence guarantees involving large language models in multi-agent systems remains largely unexplored. Systematic investigation and theoretical innovation in this area are urgently needed.

3. Method

This study proposes a multi-agent reinforcement learning method that integrates a large language model strategy guidance mechanism, aiming to improve collaboration efficiency and strategy convergence quality. The model architecture is shown in Figure 1.

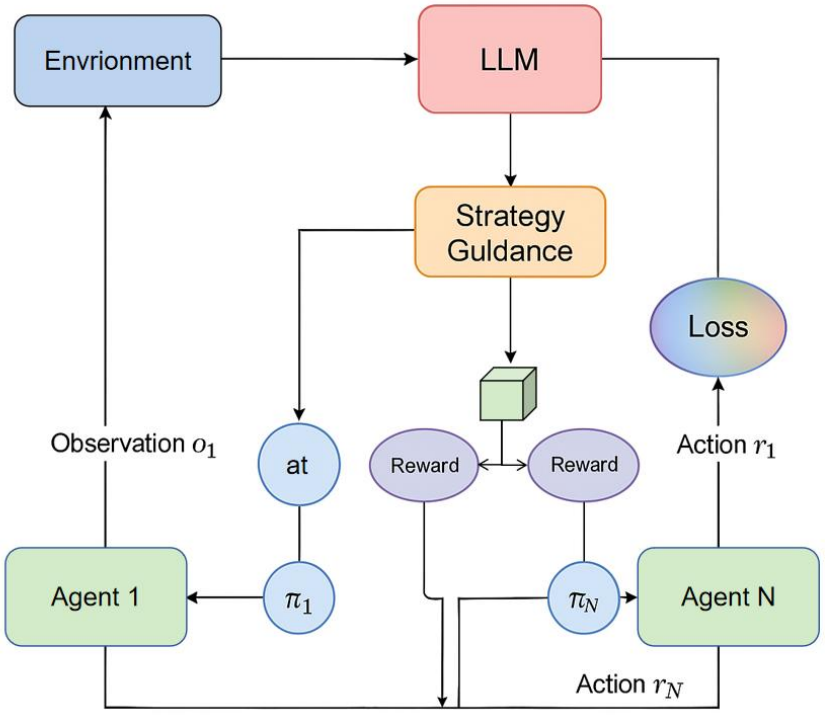

Figure 1. Overall model architecture

The model architecture diagram shows a multi-agent collaborative reinforcement learning framework that integrates a large language model guidance mechanism. The large language model generates policy

ISSN:2377-0430
Vol. 5, No. 5, 2025

recommendations based on the environment state, task description, and historical trajectory, and assists multiple agents in decision optimization through the policy guidance module. After obtaining local observations, each agent interacts with the environment based on language guidance, and finally achieves policy updates and collaborative learning through loss feedback.

Consider a collaborative system consisting of N agents, with the environment state recorded as $s \in S$, each agent $i \in \{1,2,...,N\}$ has local observable information $o_i \in O_i$, and the action space is $a_i \in A_i$. The overall strategy of the system is the joint strategy $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$, and the transfer of the environment at time step t satisfies the Markov property:

$$P(s_{t+1} \mid s_t, a_1^t, ..., a_N^t)$$

The goal of each agent is to maximize its cumulative reward $R_i = E[\sum_{t=0}^{T} \gamma^t r_i^t]$, where r_i^t is the individual reward and $\gamma \in (0,1)$ is the discount factor. Traditional methods obtain joint strategies through centralized training, but when the number of agents increases or the task dimension becomes complex, collaborative strategy learning will face the problems of dimensionality explosion and strategy oscillation.

To this end, this paper introduces a large language model $L(\cdot)$ as a strategy guidance module to provide cross-agent collaboration strategy prompts at each training stage. The language model receives the current environment state s_t, task description τ, and historical behavior trajectory h_t as input, and outputs collaboration suggestions a_i^t for each agent, namely:

$$a_i^t = L(s_t, \tau, h_t)_i$$

These suggestions do not directly replace the policy output, but participate in the individual policy optimization process in the form of "guidance items". Specifically, in the policy update, we use KL divergence to regularize the relationship between the agent's current policy $\pi_i(a_i^t \mid o_i^t)$ and the language guidance policy, and construct the following loss function:

$$L_i = L_{RL}(\pi_i) + \lambda \cdot D_{KL}(\pi_i(a_i^t \mid o_i^t) \| a''^t_i)$$

Where L_{RL} represents the standard reinforcement learning loss, such as the policy gradient or Q-learning-based objective[14], and λ is the guided strength weight parameter. This structure achieves soft alignment between the learning strategy and the language model strategy, while maintaining the ability of individual strategies to update autonomously, while introducing collaborative prior knowledge across individuals.

Furthermore, to avoid static deviations in the language model guidance process, this paper adopts a dynamic guidance mechanism based on collaborative evaluation. In each round of training, we calculate the advantage difference between the language guidance strategy and the current joint strategy execution result:

$$\Delta_i^t = Q_i(s_t, a_i'^t) - Q_i(s_t, a_i^t)$$

If $\Delta_i^t > \delta$, it means that the language suggestion is significantly better than the current strategy, and the system will increase the weight of λ, thereby strengthening the guiding role; otherwise, the influence of language prompts will be weakened. This mechanism realizes the dynamic adjustment of strategy guidance

and improves the adaptability and coordination consistency of the language model in unstable game scenarios. In addition, in order to improve training efficiency, this paper also introduces language distillation strategies at multiple stages, by integrating the past multiple rounds of language model suggestions to build a knowledge memory library, to assist the generalization of strategies in complex environments.

In summary, the method framework of this paper builds a guided regularized optimization path on the original reinforcement learning goal by integrating the strategy generation ability of the large language model with the environmental feedback learning mechanism of multiple agents. This method not only enhances the stability of collaborative strategy learning, but also significantly improves the training efficiency and system robustness. By rationally designing language guidance, regularization mechanism and dynamic adjustment strategy, this study provides a new technical path for intelligent decision-making of multiple agents in complex collaborative tasks.

4. Experiment

4.1 Datasets

This study adopts the Multi-Agent Particle Environment (MPE) as the primary experimental dataset. MPE is a lightweight simulation platform for multi-agent reinforcement learning. It is widely used to evaluate cooperative, competitive, and mixed interaction strategies among agents. The environment consists of multiple two-dimensional particle agents. These agents operate in a bounded space, receive local observations, and make decisions. MPE supports both continuous and discrete action spaces, making it suitable for constructing challenging multi-agent task scenarios.

MPE includes several sub-task environments, such as simple cooperation, entity coverage, communication, and predator-prey tasks. These sub-tasks reflect various mechanisms of collaboration and conflict among agents. In this study, we focus on two representative scenarios: Cooperative Navigation and Push Box. They are used to evaluate the spatial coordination ability and physical interaction performance of agents under language model guidance. These scenarios offer good controllability and repeatability, allowing fine-grained observation of how guided strategies influence agent behavior.

The dataset provides complete multi-agent interaction trajectories, including states, actions, rewards, and observations. This supports language modeling and policy learning based on historical experience. As MPE has been extensively validated by many mainstream multi-agent algorithms, its standardized design ensures fair comparisons. It also lays a solid foundation for transferring models to more complex environments in future research.

4.2 Experimental Results

1) Comparative experiment of large language model guidance strategy and traditional joint training strategy

This paper first gives the comparative experimental results of the large language model guidance strategy and the traditional joint training strategy. The experimental results are shown in Table 1.

Table 1: Comparative experiment of large language model guidance strategy and traditional joint training strategy

Method	Average Reward	Convergence steps	Collaboration success rate

Ours	87.6	9500	94.2%
Joint Policy Gradient Method[15]	79.3	13400	85.6%
Attention coordination mechanism[16]	81.1	12000	87.2%
Communication Enhancement Training Strategy[17]	76.5	14500	80.4%
Independent training without communication[18]	68.7	15800	73.1%

The experimental results show that the introduction of language model-guided strategies significantly outperforms traditional joint training methods in multi-agent systems. In terms of average reward, the proposed approach achieved a peak value of 87.6, demonstrating a clear advantage over other baseline methods. This indicates that strategy prompts generated by the language model effectively enhance overall agent performance in the environment. It also confirms the capability of language models to express abstract knowledge and provide actionable guidance, helping agents optimize their strategies from a global perspective.

Regarding training convergence speed, the guided strategy completed training within 9,500 steps, which is significantly fewer than those required by joint policy gradient or communication-enhanced methods. This shows that the guidance mechanism not only accelerates learning but also reduces unnecessary exploration. The acceleration mainly benefits from the behavioral priors provided by the language model, allowing agents to quickly locate high-quality policy regions in the early training phase and avoid inefficient trial-and-error loops.

In terms of cooperation success rate, the guided strategy achieved 94.2%, far exceeding the 73.1% of independent training methods. This result highlights the coordinating effect of language models in multi-agent collaboration. In contrast, independent training without communication mechanisms suffers from severe behavioral conflicts due to inconsistent strategies, leading to frequent cooperation failures. With language-guided strategies, the system can semantically align task goals and policies, fundamentally improving inter-agent collaboration.

Overall, the experiment validates the effectiveness of integrating large language models into multi-agent policy learning. The proposed method improves multiple performance metrics and shows stronger advantages in model stability and system coordination. The results support the core claim of this study: the strategy guidance mechanism provided by language models can effectively mitigate non-stationarity and policy conflicts in multi-agent reinforcement learning. It offers a practical and efficient solution for intelligent collaboration in complex task environments.

2) Experiment on the effectiveness of language model guidance under different task complexities

Furthermore, this paper presents an experiment on the effectiveness of language model guidance under different task complexities, and the experimental results are shown in Figure 2.

The experimental results show that language model-guided strategies exhibit clear advantages across different levels of task complexity, with particularly strong performance in more complex tasks. In simple tasks, the guided strategy achieved a success rate of 95.2%, nearly 8 percentage points higher than baseline

methods. This indicates that even in low-complexity scenarios, prior strategies provided by language models can still improve cooperation efficiency and execution accuracy.

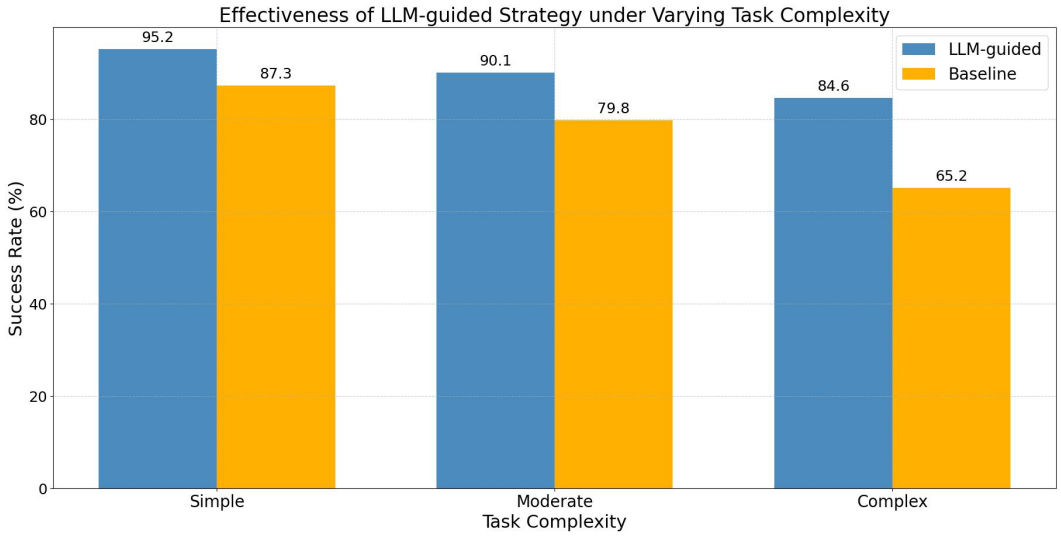

Figure 2. Effect of Multi-round Language Distillation on Agent Generalization

As task complexity increases, the performance of baseline methods declines more significantly. In contrast, the language model-guided strategy maintains a high success rate, reaching 90.1% in medium-complexity tasks and 84.6% in high-complexity tasks. This suggests that language models have a certain degree of generalization ability. They can effectively perceive global environmental information and reason about cooperative strategies, helping agents deal with conflicts and exploration challenges arising from increased task dimensions.

Overall, the experiment confirms the robustness and adaptability of language-guided strategies in complex multi-agent tasks. Compared to traditional methods that rely on slow convergence through environment feedback, the language model serves as an external knowledge source. It offers structured policy suggestions, significantly shortening the decision optimization process and improving task completion and cooperation success rates.

3) Strategy Stability Evaluation under Multi-agent Scale-up

Furthermore, this paper gives an evaluation of the strategy stability under the multi-agent scale expansion, and the experimental results are shown in Figure 3.

The experimental results show that as the number of agents increases, the overall strategy stability of the system tends to decline. However, the strategy guided by the language model demonstrates higher stability scores across all scales. During the scaling process from 2 to 12 agents, the guided strategy consistently maintains high stability scores and exhibits smooth degradation. This indicates strong adaptability and robustness in cooperative settings.

In contrast, baseline methods show a significant drop in stability as the number of agents increases, especially when the number exceeds 8. This gap mainly results from the tendency of traditional methods to suffer from strategy interference and coordination conflicts in large-scale multi-agent scenarios. They lack a unified coordination mechanism. In comparison, the language model-guided strategy provides each agent with globally consistent strategy prompts. This helps mitigate instability caused by inter-agent competition at the semantic level.

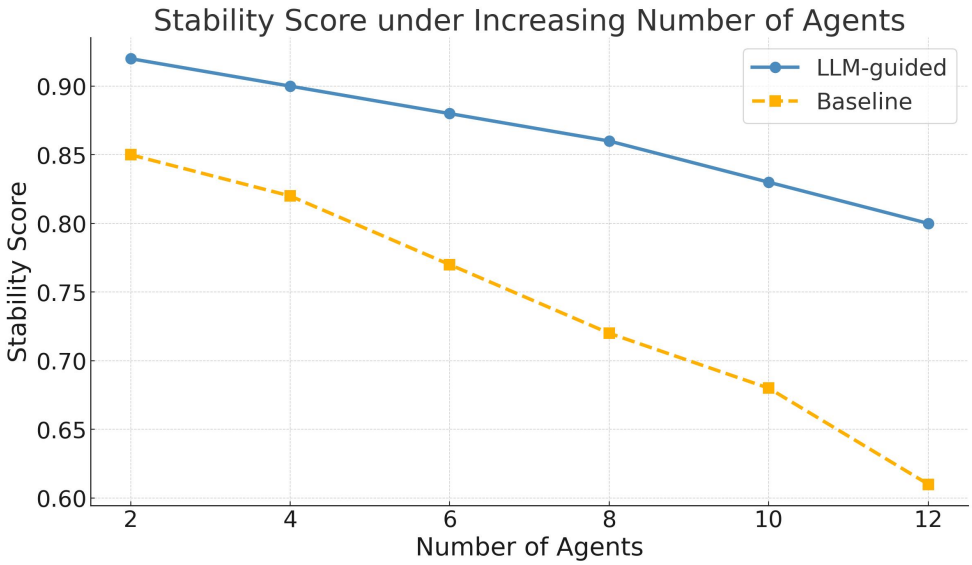

Figure 3. Stability Score under Increasing Number of Agents

In summary, the experiment further confirms the coordinating capability of language models in multi-agent collaboration systems. The approach shows superior performance in terms of scalability. By introducing the language model as an auxiliary module for policy generation, it becomes possible to maintain consistent and stable strategy convergence even in high-density agent systems. This provides both theoretical and practical support for the scalable application of multi-agent reinforcement learning in complex tasks.

4) Analysis of the influence of guidance intensity adjustment parameters on individual performance

Furthermore, the influence of the guidance intensity adjustment parameters on the individual performance is analyzed. The actual results are shown in Figure 4.

The experimental results indicate that the guidance strength parameter has a significant impact on individual performance. As the guidance strength increases, performance first improves and then slightly declines. When the parameter is set to 0.5, the performance reaches the highest score of 88.6%. This suggests that a moderate level of guidance from the language model provides the most effective policy prompts. It helps optimize the decision-making of individual agents and maximizes cooperative gains.

However, when the guidance strength increases further to 0.7 and 0.9, performance begins to decline. This indicates that excessive guidance may limit the flexibility of individual policy optimization. It reduces exploration and weakens the agents' ability to adapt to dynamic environments. Similarly, when the guidance strength is too low (e.g., 0.1), the benefits of the language model cannot be fully utilized. Agents struggle to access global policy insights efficiently, which results in reduced overall performance.

In summary, the results clearly demonstrate the effectiveness and adaptability of language model-guided strategies in multi-agent reinforcement learning. They also highlight the importance of carefully tuning the guidance strength to achieve optimal system performance. This finding further validates the design of the guidance strength control mechanism in this study and provides a valuable reference for enabling stable and efficient multi-agent collaboration in complex tasks.

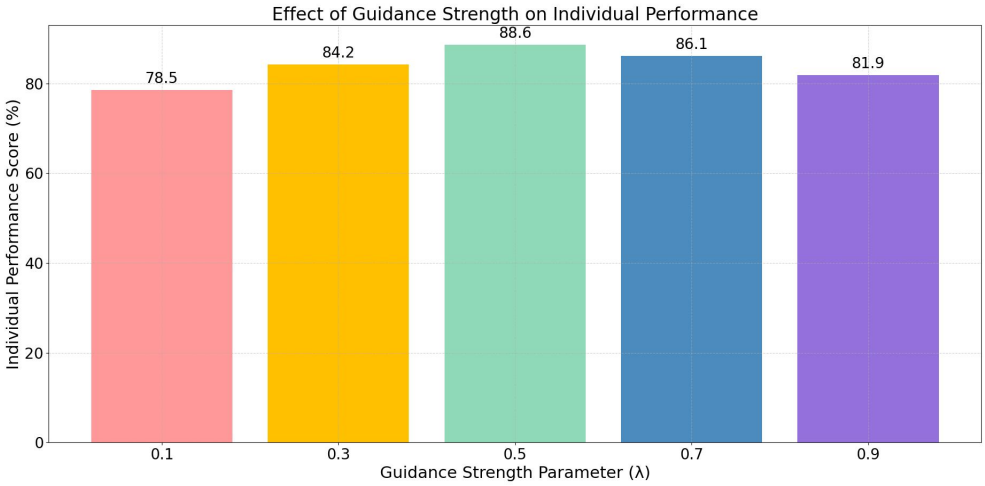

Figure 4. Effect of Guidance Strength on Individual Performance

5) Analysis of the impact of dynamic guidance mechanism on training convergence

Finally, an analysis of the impact of the dynamic guidance mechanism on training convergence is given, and the experimental results are shown in Figure 5.

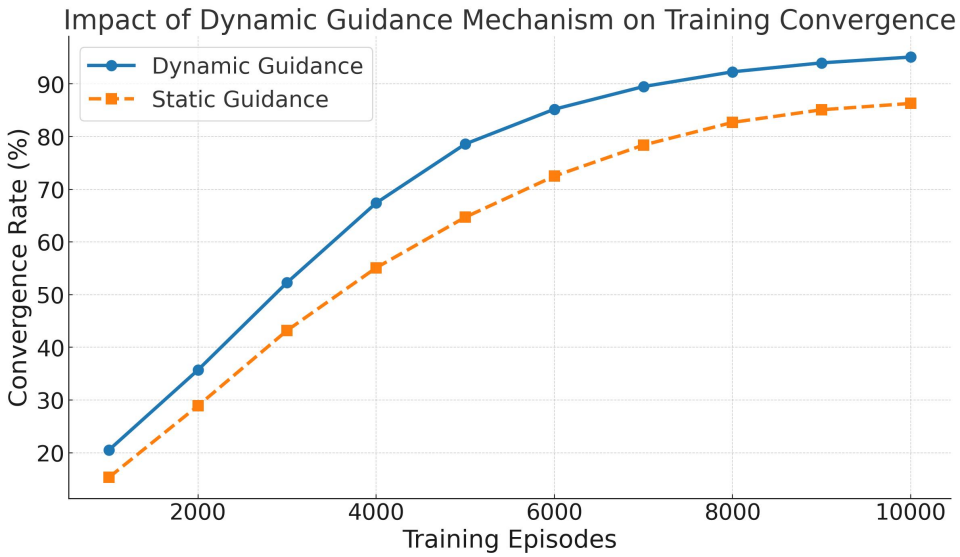

Figure 5. Impact of Dynamic Guidance Mechanism on Training Convergence

The experimental results show that the dynamic guidance mechanism significantly improves convergence efficiency compared to the static guidance mechanism. Throughout the training cycle, the dynamic mechanism consistently achieves a faster convergence rate. This advantage is especially prominent in the early training stage (between 2,000 and 6,000 episodes). It suggests that dynamically adjusting the guidance strength of the language model helps agents quickly locate better policy regions and greatly accelerates policy optimization.

As training progresses, the gap in convergence between the two mechanisms narrows. However, by 10,000 training episodes, the dynamic mechanism still reaches a convergence rate of 95.1%, which is notably higher than the 86.3% achieved by the static mechanism. This indicates that the dynamic approach not only speeds up early learning but also enhances stability and accuracy in later-stage policy refinement. It helps reduce policy oscillation and ineffective exploration.

Overall, the experiment fully verifies the effectiveness of the proposed dynamic guidance mechanism in improving training efficiency in multi-agent reinforcement learning. By continuously adjusting the guidance strength from the language model, the dynamic mechanism maximizes the model's advantages in policy generation and knowledge-based decision support. It provides crucial support for achieving faster and more stable multi-agent collaboration training.

5. Conclusion

This paper proposes a multi-agent cooperative reinforcement learning method guided by large language model strategies. The aim is to address key challenges in multi-agent reinforcement learning, such as policy instability and low training efficiency. Experimental results demonstrate that the proposed guidance strategy significantly improves cooperation performance in multi-agent systems. It also accelerates convergence and consistently outperforms traditional methods under varying task complexities, agent scales, and guidance strength settings. These findings confirm that language models, as policy guidance tools, can successfully provide high-quality strategy prompts, reduce exploration costs, and enhance coordination efficiency among agents.

Specifically, the introduction of a dynamic guidance mechanism allows the model to adjust guidance strength in real time based on environmental and task feedback. This improves both the flexibility and stability of policy learning. In addition, the paper investigates the impact of guidance strength on individual performance and finds that appropriate guidance levels are crucial for effective policy optimization. The experimental results fully validate the feasibility and superiority of this method in improving training efficiency and agent performance. It provides a new approach and technical pathway for deploying multi-agent cooperation in complex environments.

Despite the promising results achieved in this study, several directions remain for future exploration. For instance, how to more efficiently integrate language models with reinforcement learning frameworks to reduce computational overhead while enhancing policy generalization remains an open question. Furthermore, exploring the adaptability of language-guided mechanisms in heterogeneous multi-agent systems is also a key topic for future work. With the continuous development of relevant technologies, the increasing scale and expressive power of large language models offer great potential to further advance the intelligence of multi-agent systems. Looking ahead, language model-based policy guidance mechanisms hold promise not only for theoretical research and simulation environments but also for real-world applications in industrial, military, and societal domains. They can help build more efficient and stable multi-agent decision-making systems. At the same time, integrating multimodal inputs — such as vision and sound — into the language-guided framework will further expand the perception and decision-making capabilities of multi-agent systems. This will significantly enhance their adaptability in unknown and complex environments, making it one of the key directions for future reinforcement learning research.

References

[1] Park, Chanwoo, et al. "Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning." arXiv preprint arXiv:2502.18439 (2025).

[2] Guo, Taicheng, et al. "Large language model based multi-agents: A survey of progress and challenges." arXiv preprint arXiv:2402.01680 (2024).

[3] Wen, Muning, et al. "Multi-agent reinforcement learning is a sequence modeling problem." Advances in Neural Information Processing Systems 35 (2022): 16509-16521.

[4] Wang, Ziyan, et al. "Safe multi-agent reinforcement learning with natural language constraints." arXiv preprint arXiv:2405.20018 (2024).

[5] Sun, Chuanneng, Songjun Huang, and Dario Pompili. "Llm-based multi-agent reinforcement learning: Current and future directions." arXiv preprint arXiv:2405.11106 (2024).

[6] Qian, Chen, et al. "Scaling large-language-model-based multi-agent collaboration." arXiv preprint arXiv:2406.07155 (2024).

[7] Sarkar, Bidipta, et al. "Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning." arXiv preprint arXiv:2502.06060 (2025).

[8] Slumbers, Oliver, et al. "Leveraging large language models for optimised coordination in textual multi-agent reinforcement learning." (2023).

[9] Li, Huao, et al. "Theory of mind for multi-agent collaboration via large language models." arXiv preprint arXiv:2310.10701 (2023).

[10] Jiang, Feibo, et al. "Large language model enhanced multi-agent systems for 6G communications." IEEE Wireless Communications (2024).

[11] Alsadat, Shayan Meshkat, and Zhe Xu. "Multi-Agent Reinforcement Learning in Non-Cooperative Stochastic Games Using Large Language Models." IEEE Control Systems Letters (2024).

[12] Yu, Jiapeng, et al. "Co-Learning: Code Learning for Multi-Agent Reinforcement Collaborative Framework with Conversational Natural Language Interfaces." arXiv preprint arXiv:2409.00985 (2024).

[13] Liu, Zhihao, et al. "Knowing What Not to Do: Leverage Language Model Insights for Action Space Pruning in Multi-agent Reinforcement Learning." arXiv preprint arXiv:2405.16854 (2024).

[14] Shen, Xiao-Ning, et al. "A Q-learning-based memetic algorithm for multi-objective dynamic software project scheduling." Information Sciences 428 (2018): 1-29.

[15] Zhao, Li-yang, et al. "Multi-agent cooperation policy gradient method based on enhanced exploration for cooperative tasks." International Journal of Machine Learning and Cybernetics 15.4 (2024): 1431-1452.

[16] Hu, Kai, et al. "An overview: Attention mechanisms in multi-agent reinforcement learning." Neurocomputing (2024): 128015.

[17] Zhu, Changxi, Mehdi Dastani, and Shihan Wang. "A survey of multi-agent deep reinforcement learning with communication." Autonomous Agents and Multi-Agent Systems 38.1 (2024): 4.

[18] Tan, Ming. "Multi-agent reinforcement learning: Independent vs. cooperative agents." Proceedings of the tenth international conference on machine learning. 1993.