Stackelberg Decision Transformer for Asynchronous Action Coordination in Multi-Agent Reinforcement Learning

Bin Zhang
2024/01/03

中国科学院大学
University of Chinese Academy of Sciences
CASIA
中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Game Theory

Equilibrium signifies that in a multiparty game, all players have adopted the optimal strategy and none can improve their performance by altering their own strategy.

Nash Equilibrium

$$v^j(s;\boldsymbol{\pi}_*) = v^j(s;\pi_*^j, \boldsymbol{\pi}_*^{-j}) \geq v^j(s;\pi^j, \boldsymbol{\pi}_*^{-j})$$

- Nash Q-Learning;
- Mean Field Q-learning;
- HATRPO

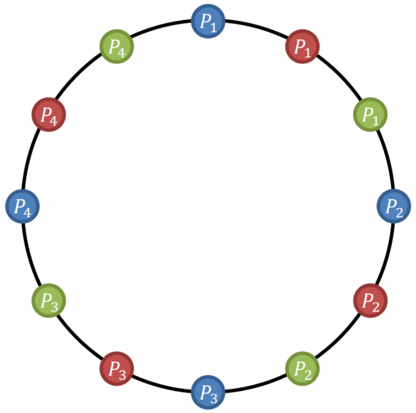
Stackelberg Equilibrium

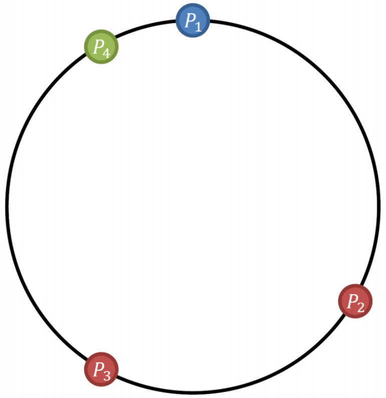
$$V^1_{\pi^{1*},\pi^{2*}}(s) \geq V^1_{\pi^1,\pi^{2*}}(s),$$
$$V^2_{\pi^1,\pi^{2*}}(s,a^1) \geq V^2_{\pi^1,\pi^2}(s,a^1).$$

- Asymmetric Q-learning;
- Bi-level Actor Critic

a^1 \ a^2	a^2_1	a^2_2	a^2_3
a^1_1	k	0	**10**
a^1_2	0	**2**	0
a^1_3	**8**	0	k

a^1 \ a^2	a^2_1	a^2_2	a^2_3
a^1_1	**0, 5**	-10,-5	-8, 4
a^1_2	-5,-10	**-5, 0**	-15,-5
a^1_3	**5, 0**	-10,-5	-10, 5





University of Chinese Academy of Sciences

Motivation

Stackelberg Game

$$\max_{\pi^1\in\Pi^1}\{\mathcal{J}^1(\pi^1,\pi^2)|\pi^2\in\arg\max_{\pi^{2'}\in\Pi^2}\mathcal{J}^2(\pi^1,\pi^{2'})\},$$

$$\max_{\pi^2\in\Pi^2}\mathcal{J}^2(\pi^1,\pi^2),$$

Stackelberg Equilibrium

$$V^1_{\pi^{1*},\pi^{2*}}(s)\geq V^1_{\pi^1,\pi^{2*}}(s),$$

$$V^2_{\pi^1,\pi^{2*}}(s,a^1)\geq V^2_{\pi^1,\pi^2}(s,a^1).$$

Stackelberg Equilibrium

- The paradigm of sequential decision-making is conceptually defined from the perspective of game theory.
- applicable to both cooperative and non-cooperative games.
- surpasses Nash equilibrium in terms of equilibrium determinacy and Pareto optimality.

When SE encounters MARL, we aim to address the following challenges:

- How to make a reinforcement learning algorithm converge to the Stackelberg equilibrium strategy?
- How to converge to SE policies that require agents act sequentially under the MG framework where agents act simultaneously?
- How to extend the method to scenarios with more than two agents (n > 2)?

STMG

N-level optimization

$$\max_{\pi^i \in \Pi^i} \{\mathcal{J}^i(\pi^{1:i-1}, \pi^i) | \pi^j \in \arg\max_{\pi^{j'} \in \Pi^j} \mathcal{J}^j(\pi^{1:j'-1}, \pi^{j'})\},$$

$$\max_{\pi^j \in \Pi^j} \mathcal{J}^j(\pi^{1:j-1}, \pi^j), \quad i \in [1:n], j \in [i+1:n]$$

Spatio-Temporal Sequential Markov Game (STMG)

Definition 1. *STMG can be formalized as a tuple $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, P, \{\nabla^i\}_{i\in\mathcal{I}}, \gamma, \{o^i\}_{i\in\mathcal{I}} \rangle$. In addition to the MG defined in 2.1, STMG add the term o^i, which denotes the action order of agent i and $\mathcal{O} = (o^1, ..., o^n)$ represents all agents' action order, indicating the priority/importance of agents at the decision-making stage.*

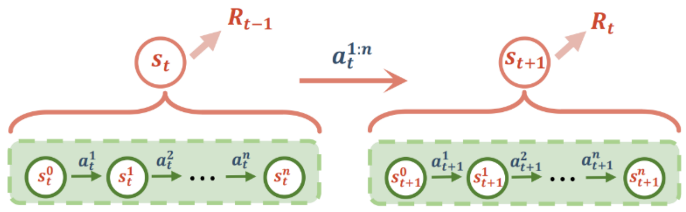


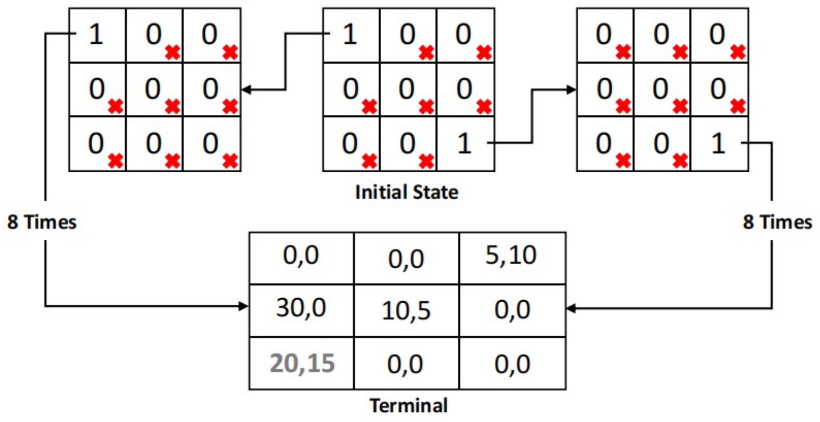
Figure 1: The STMG state transition procedure. It is an extensive game version of MG, which specifies the decision-making sequence of agents simultaneously.

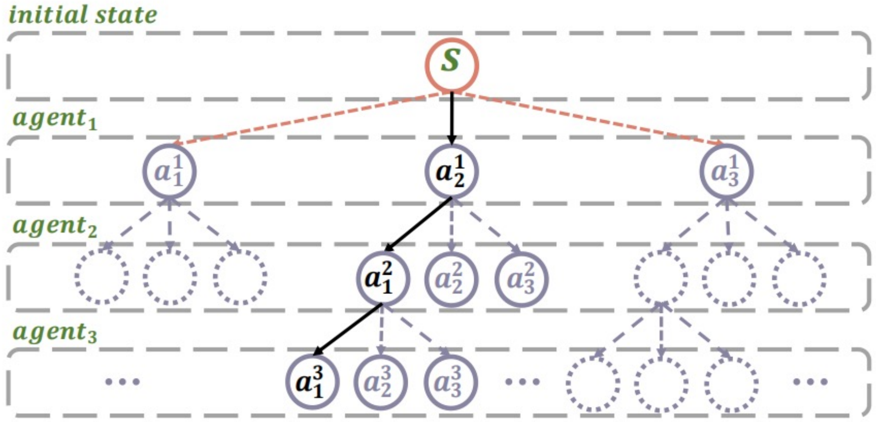
Compared with MG, STMG assumes the form of a sequence decision in both temporal and spatial domains. Agents with a higher priority have greater initiative, whereas agents with a lower priority are required to respond to the actions of those with higher priority.

$$Q_{\boldsymbol{\pi}}^{h^i}(s, a^{h^1:h^{i-1}}, a^{h^i}) = \mathbb{E}_{s\sim\rho, \boldsymbol{a}\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r_t^{h^i}(s_t, \boldsymbol{a_t}) \mid \mathrm{s}_0 = s, \boldsymbol{a}_0^{h^1:h^i} = \boldsymbol{a}^{h^1:h^i}\right],$$

$$V_{\boldsymbol{\pi}}^{i}(s, a^{h^1:h^{i-1}}) = \sum_{a^{h^i} \in \mathcal{A}^{h^i}} \pi^i(a^{h^i} | s, a^{h^1:h^{i-1}}) Q_{\boldsymbol{\pi}}^{h^i}(s, a^{h^1:h^{i-1}}, a^{h^i}).$$

Commencing with a Toy Example





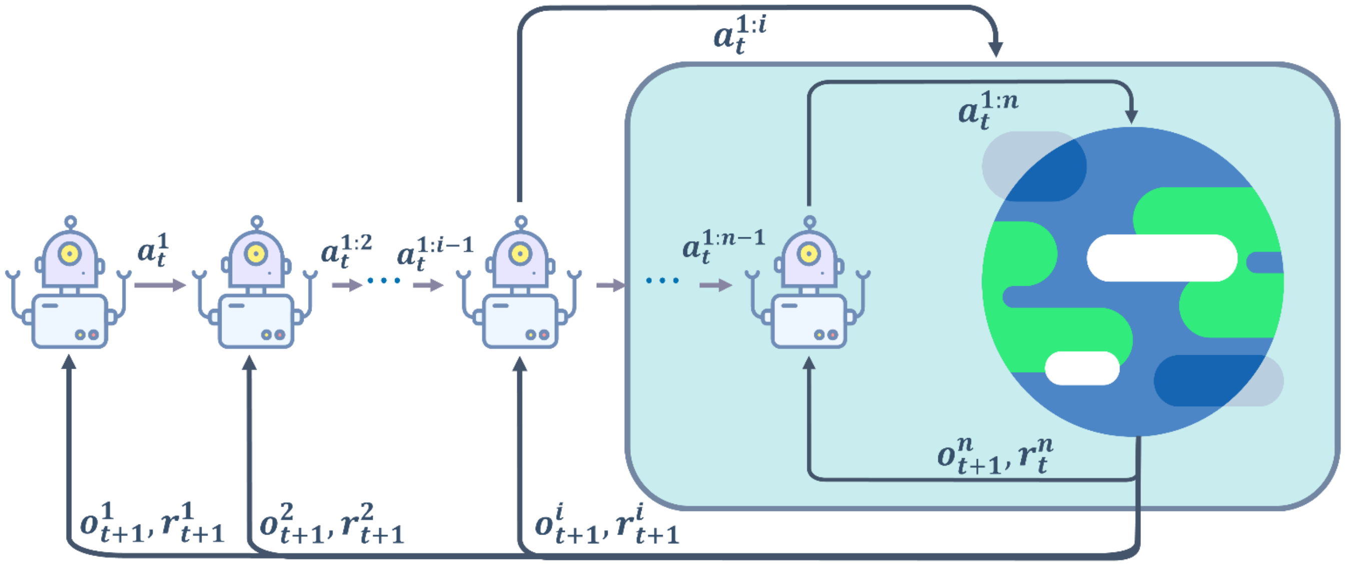
- When the leader commits to taking action, the ideal space for followers to take action is constrained.
- In the final state, all three joint actions $(a_1^1\ ,\ a_3^2), (a_2^1\ ,\ a_2^2)$ and $(a_3^1\ ,\ a_1^2)$, are Nash equilibrium (NE) points. However, only the point$(a_1^1\ ,\ a_3^2)$is the unique socially efficient (SE) point and also the global optimum.

Algorithm design requirements :

- All agents possess accurate perceptual awareness of the current state.
- The environmental state and the leader's decision information must be taken into account during policy evaluation and execution.

中国科学院大学 University of Chinese Academy of Sciences

Heuristic Stackelberg Decision Mechanism for MARL



- Followers directly receive decision information from higher-level agents, and the agent's policy gradient is updated towards the optimal response to the higher-level agent, resulting in an approximate solution to the inner optimization problem.
- Leaders interact with the environment and perceive the reaction of the inferior agents.

Under the RL training paradigm, all agents possess the capability to maximize their individual utility in accordance with current conditions, thereby naturally achieving corresponding equilibrium.

University of Chinese Academy of Sciences

STEP

Implementation :

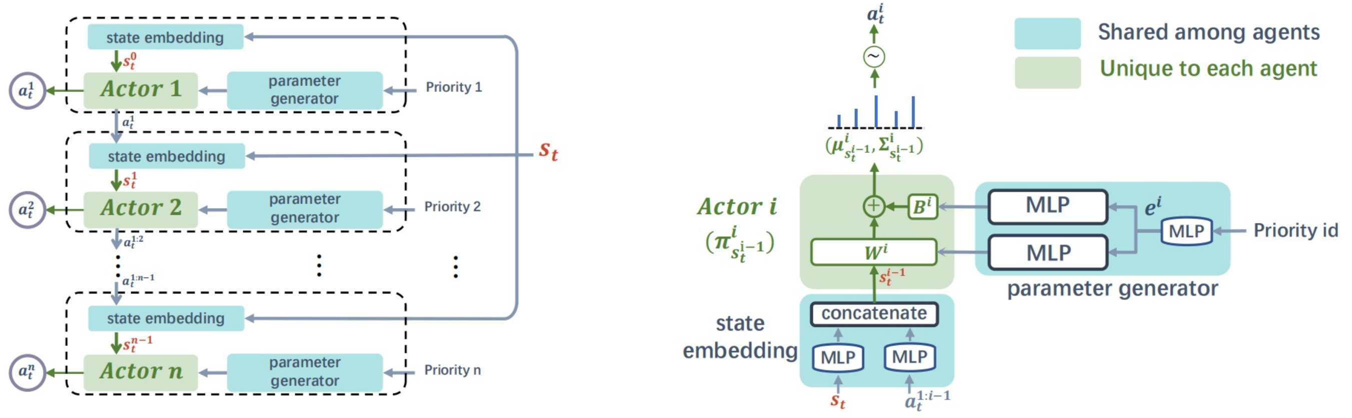


Figure 3: The overall architecture of STEP. *Left*: The workflow of STEP for a comprehensive decision in a time step. Agents base their decisions on the current situation s_t, their self-positioning $Priority$ ID, and the prerequisite actions $a_t^{1:i-1}$ of superior agents. *Right*: The structure of N-level policy model. It allows for the implementation of heterogeneous policies under parameter sharing and the Stackelberg equilibrium policies under symmetric conditions.

Limitations:

- Focus on CTDE/ATSE paradigm;
- Only applicable to situations where a shared global state is present.
- Sequential updates result in a significant increase in training costs as the number of agents grows.

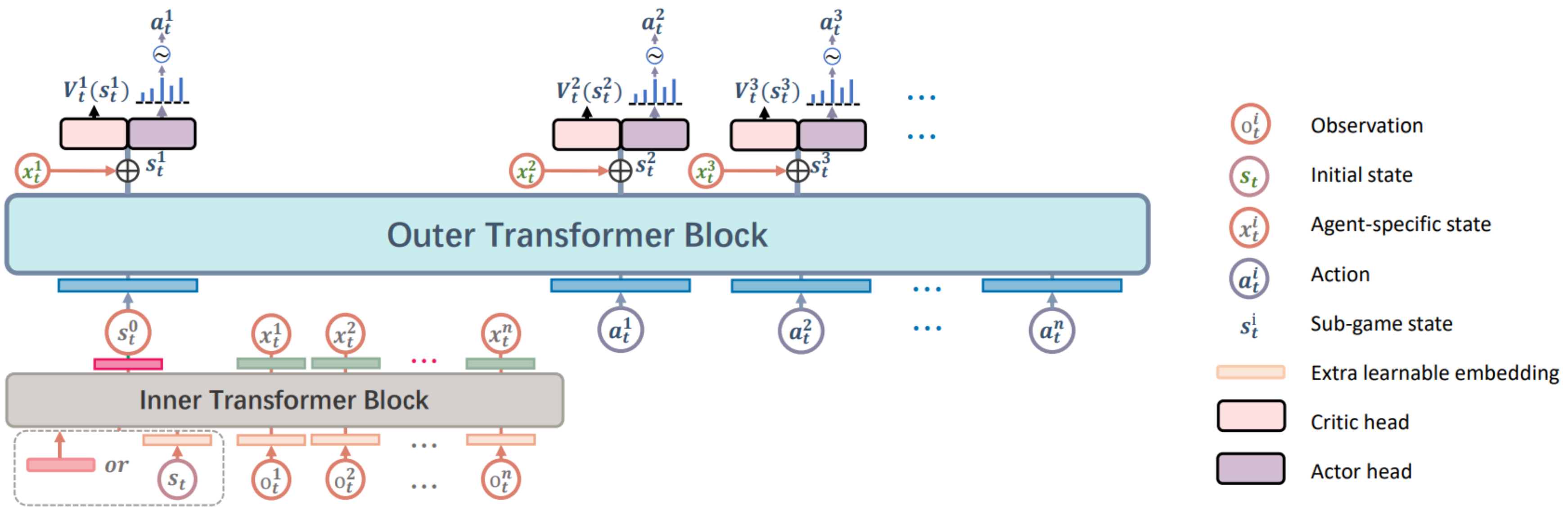
What is a better solution?

Causal Transformer!

University of Chinese Academy of Sciences

Stackelberg Decision Transformer

The seamless alignment between the hierarchical decision-making structure of SG and the modeling approach of autoregressive sequence models.



ITB

$$e'_{\ell_j,t} = \mathrm{MHSA}(\mathrm{LN}(e_{\ell_{j-1},t})) + e_{\ell_{j-1},t},$$
$$e_{\ell_j,t} = \mathrm{MLP}(\mathrm{LN}(e'_{\ell_j,t})) + e'_{\ell_j,t}.$$
$$Y_t^{ITB} = \mathrm{MLP}(e_{L,t}) = [s_t^0, x_t^1, ..., x_t^n],$$

OTB

$$z'_{\ell_j,t} = \mathrm{MMHSA}(\mathrm{LN}(z_{\ell_{j-1},t})) + z_{\ell_{j-1},t},$$
$$z_{\ell_j,t} = \mathrm{MLP}(\mathrm{LN}(z'_{\ell_j,t})) + z'_{\ell_j,t},$$
$$Y_t^{OTB} = \mathrm{MLP}(z_{L,t}).$$

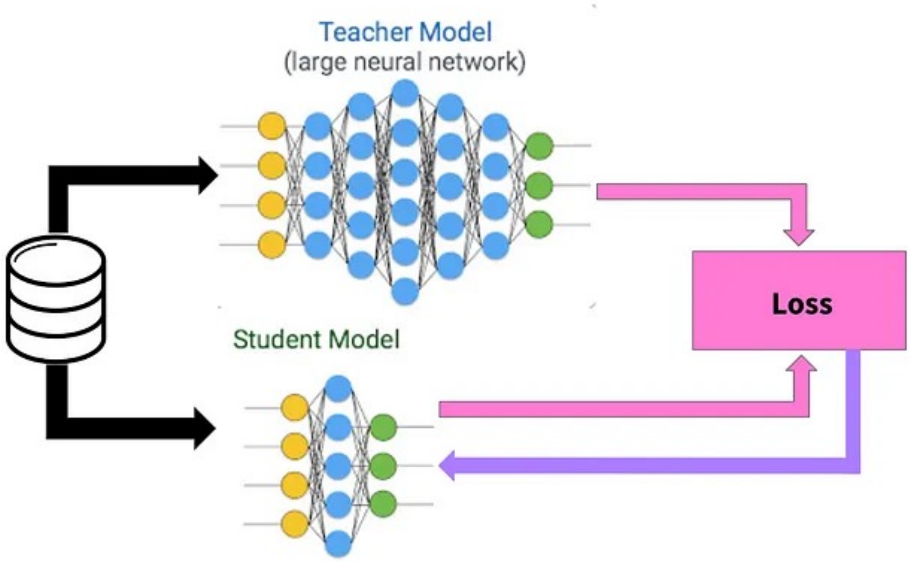
University of Chinese Academy of Sciences

Stackelberg Decision Transformer

Scalability for Decentralized Execution Systems—Knowledge Distillation

Forward propagation in the Transformer-based STEER teacher network

Backpropagation in multiple MLP-based student networks



$$L_{\text{student}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log\left(\pi_{\text{student}}\left(\bar{a} \mid o\right) - \log\left(\pi_{\text{STEER}}\left(\bar{a} \mid s\right)\right)^{2}\right.\right.}$$
$$-\eta S\left(\pi_{\text{student}}\left(a \mid o\right)\right)$$

Evaluation

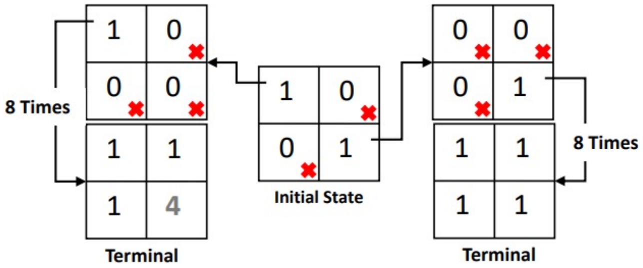
Finding SE Solutions

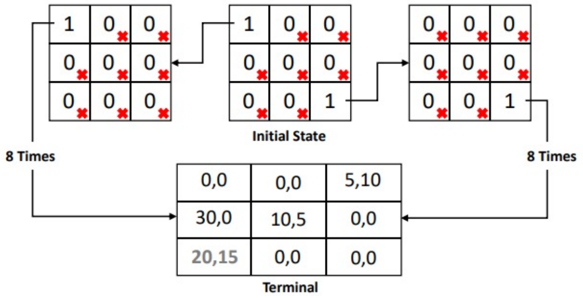
Mixing

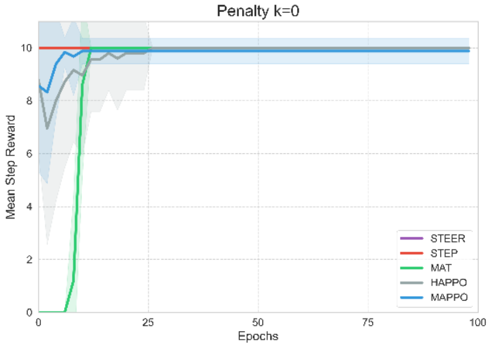
a^1 \ a^2	a^2_1	a^2_2	a^2_3
a^1_1	0, 5	-10,-5	-8,4
a^1_2	-5,-10	-5, 0	-15,-5
a^1_3	5, 0	-10,-5	-10,5

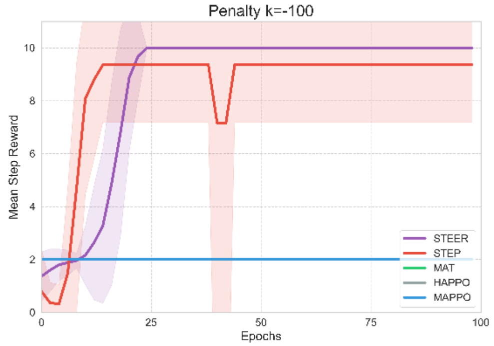
Penalty

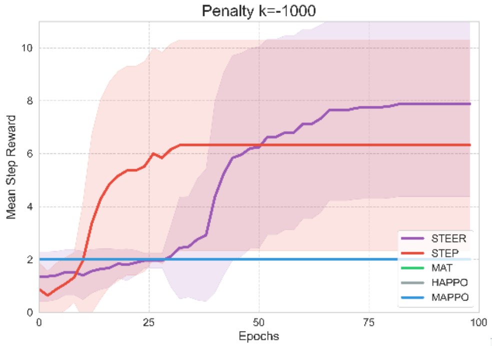
a^1 \ a^2	a^2_1	a^2_2	a^2_3
a^1_1	k	0	10
a^1_2	0	2	0
a^1_3	8	0	k



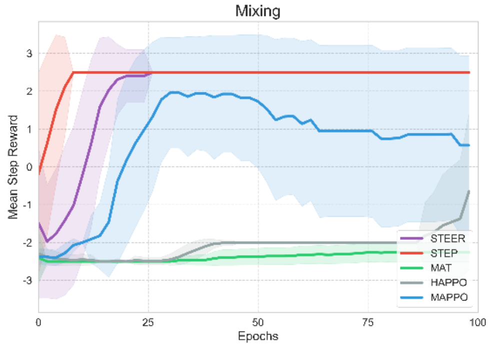


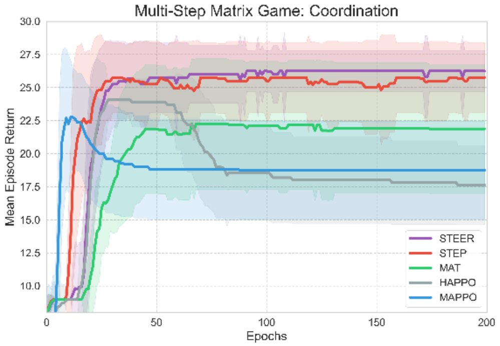


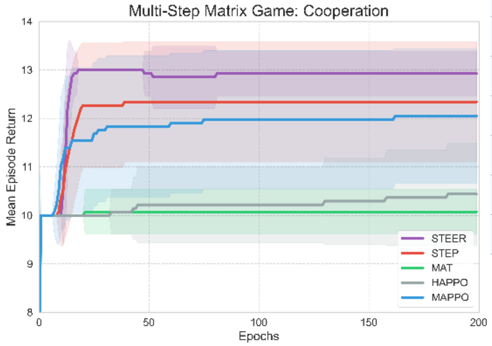




		STEER	STEP	MAT	HAPPO	MAPPO
	k=0	**10.0(0)**	**10.0(0)**	**10.0(0)**	**10.0(0)**	9.90(0.43)
Penalty	k=-100	**10.0(0)**	9.44(2.04)	2.0(0)	2.0(0)	2.0(0)
	k=-1000	**8.0(3.39)**	5.52(3.97)	2.0(0)	2.0(0)	2.0(0)
Mixing		**2.5(0)**	**2.5(0)**	-2.68(0.25)	-0.74(2.33)	0.72(2.33)
coordination		**26.0(2.42)**	25.9(2.37)	21.32(4.94)	17.62(2.94)	19.17(4.05)
cooperation		**12.88(0.58)**	12.69(0.89)	10.15(0.65)	10.57(1.18)	11.95(1.43)





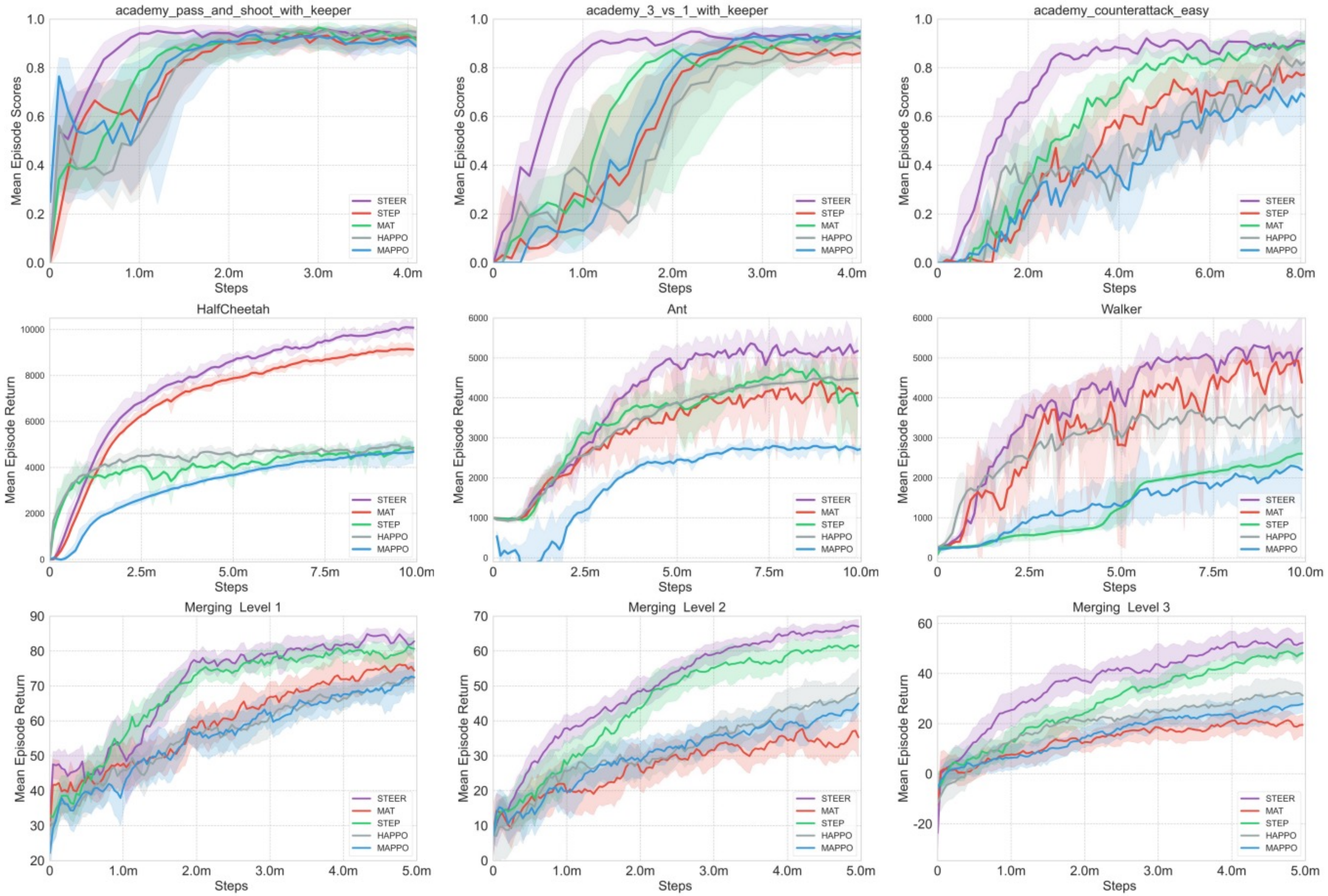


	Penalty			Mixing	coordination	cooperation
	k=0	k=-100	k=-1000			
STEER	**100**	**100**	**72**	**100**	**95**	**96**
STEP	**100**	93	44	**100**	94	90
MAT	**100**	0	0	0	46	5
HAPPO	**100**	0	0	28	6	19
MAPPO	95	0	0	63	14	65

中国科学院大学
University of Chinese Academy of Sciences

Evaluation

Performance in Complex Scenarios

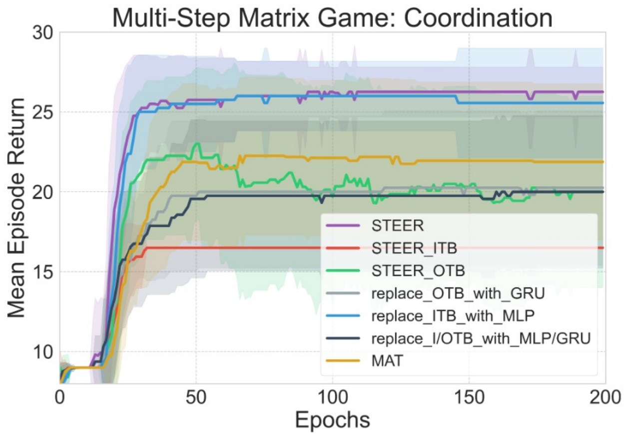


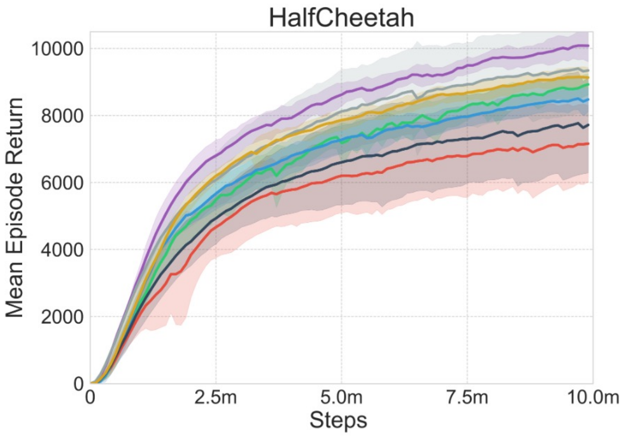
University of Chinese Academy of Sciences

Evaluation

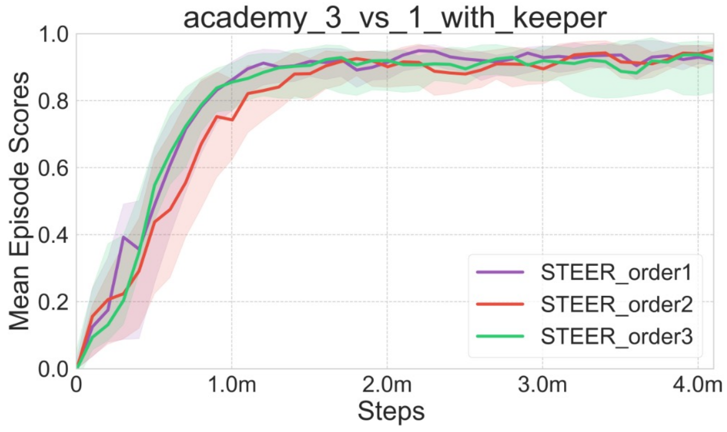
Ablation Studies

ITB & OTB





Priority Assignment



Decentralized Execution

	academy_pass_and_shoot_with_keeper	academy_3_vs_1_with_keeper	academy_counterattack_easy
STEER	0.9339(0.0358)	0.9636(0.0375)	0.9176(0.0815)
Decentralized Sudent Network	0.9426(0.0143)	0.9417(0.0203)	0.9025(0.0190)

中国科学院大学 University of Chinese Academy of Sciences

Controlling Large Language Model-based Agents for Large-Scale Decision-Making: An Actor-Critic Approach

Bin Zhang

2024/01/03

Existing Work

1、Natural Language Processing

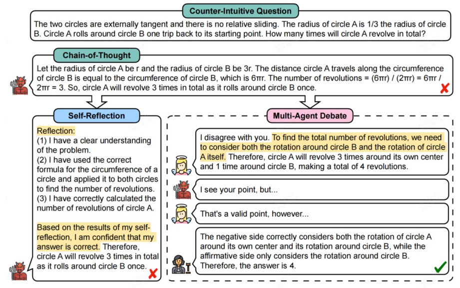


Figure 2: Framework of Multi-Agent Debate. Here we designate the devil ( ) as the affirmative side while the angel ( ) as the negative side. We want the angel to correct the devil's mistakes.

Figure 2: Qualitative examples. The correct choices are underlined.

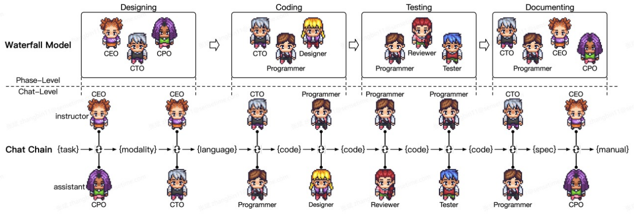


Figure 2: The proposed architecture of CHATDEV consists of phase-level and chat-level components. At the phase level, the waterfall model is used to break down the software development process into four sequential phases. At the chat level, each phase is further divided into atomic chats. These atomic chats involve task-oriented role-playing between two agents, promoting collaborative communication. The communication follows an instruction-following style, where agents interact to accomplish a specific subtask within each chat.

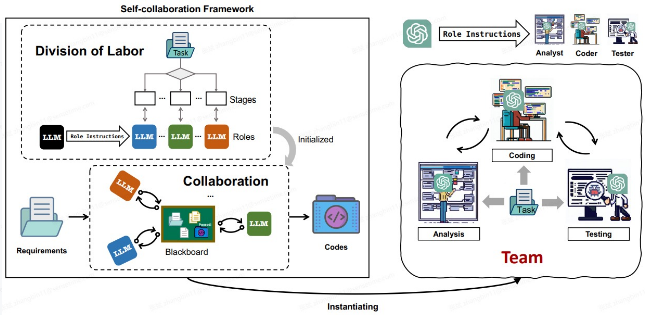


Figure 2: Self-collaboration framework for code generation and its instance.

2、Decision Making

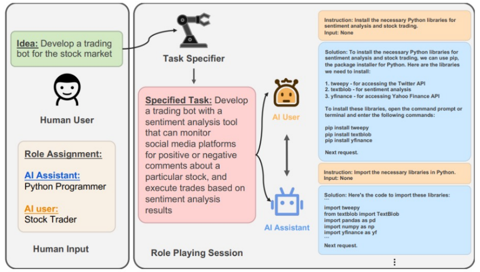


Figure 1: **Role-Playing Framework.** Our role-playing setup starts with the human user having an idea they want to implement, *e.g.* develop a trading bot for the stock market. The roles involved in this task would be an AI assistant agent who is a python programmer and an AI user agent who is a stock trader. The task is made more specific using our task specifier agent, leading to a well-defined task for the assistant to solve. The AI user and AI assistant collaboratively communicate by chatting with each other in an instruction-following fashion to solve the specified task.

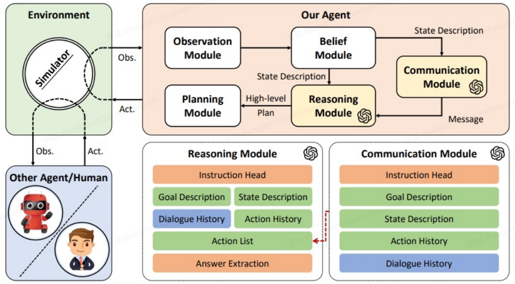
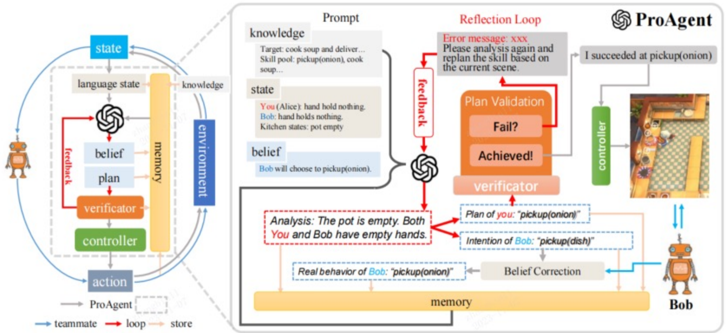


Figure 2: An overview of our framework, consisting of five modules: observation, belief, communication, reasoning, and planning, where the Communication Module and the Reasoning Module leverage Large Language Models to generate messages and decide on high-level plans. Here we also show the overall prompt design for leveraging LLMs to serve as these two modules. More design details can be found in Appendix A.

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.



Motivation

1. As the number of agents increases, the joint action space grows exponentially.
2. The limitations of LLMs themselves, such as the issue of hallucinations, can affect the reliability of decision-making.
3. Effectively managing tokens or communication resources poses a significant challenge in large-scale scenarios involving LLM-based agents.

Type	Method	Target	Role	Agents Num.
Muti-Agent Debate	Debate (Du et al.)	Task Solver	2 debaters	2
	MAD (Liang et al.)		1 judge + 2 debaters	3
	ChatEval (Chan et al.)		multi debaters	5
Role Playing	CAMEL (Li et al.)	Task Solver	1 assistant + 1 user	2
	AgentVerse (Chen et al.)		1 role assigner + 2-4 experts + 1 evaluater	6
	Proagent (Zhang et al.)		2 cooks	2
	LLaMAC (ours)		**3 critic + 1-50 actors**	**50**
	Generative Agents (Park et al.)	Community Simulator	25 agents	25
	Werewolf Agents (Xu et al.)		7 players	7
	ReCon (Wang et al.)		6 players	6

Mothod

1、 Multi-agent Actor-Critic architecture

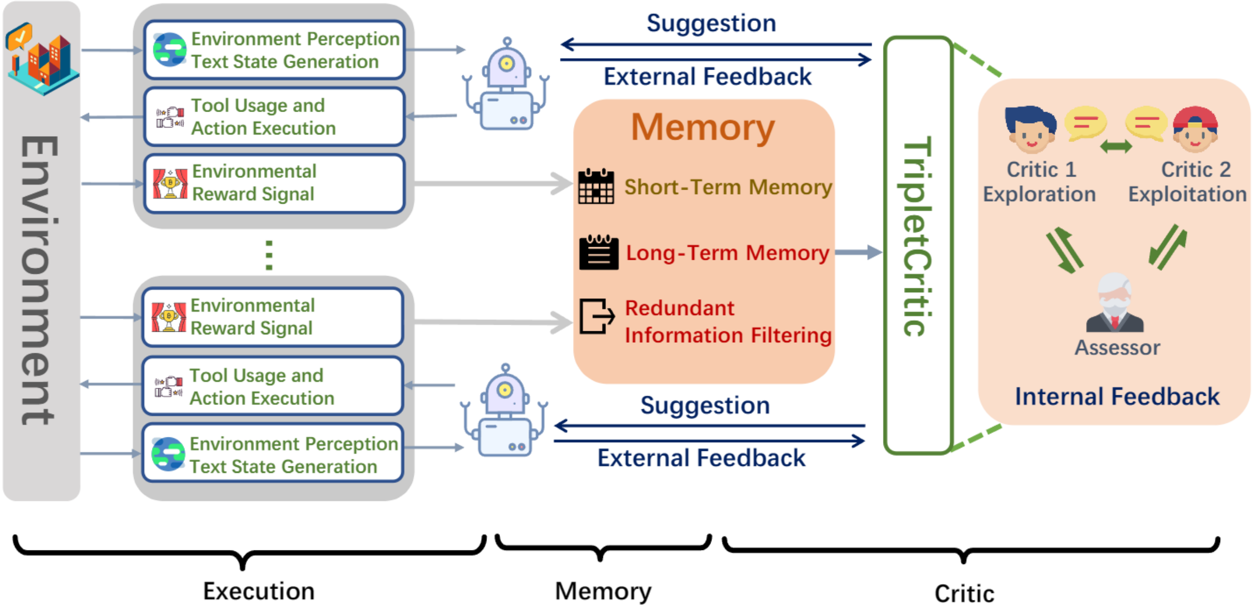
a. critic: Central Coordinator, Balancing Exploration and Exploitation, Task Allocation for Actors based on Memory Information

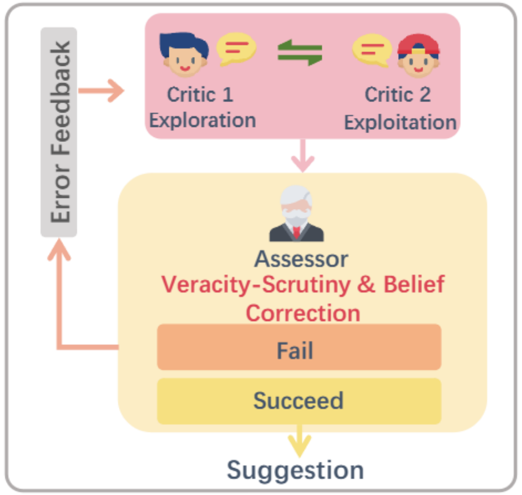
b. actor: Interaction with the environment, external feedback

2、 Large-scale Multi-Agent System Decision Making

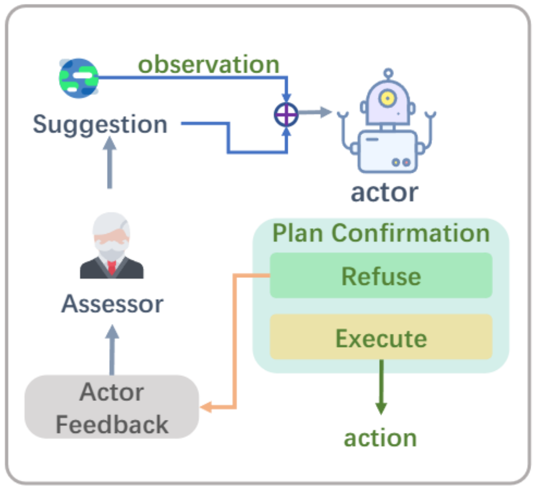
a. Comprehensive Feedback Mechanism

b. Low Access Cost





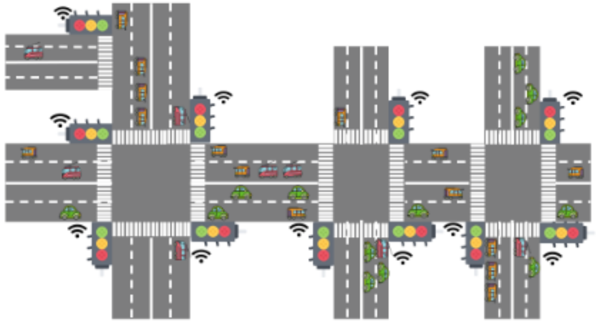
Internal Feedback

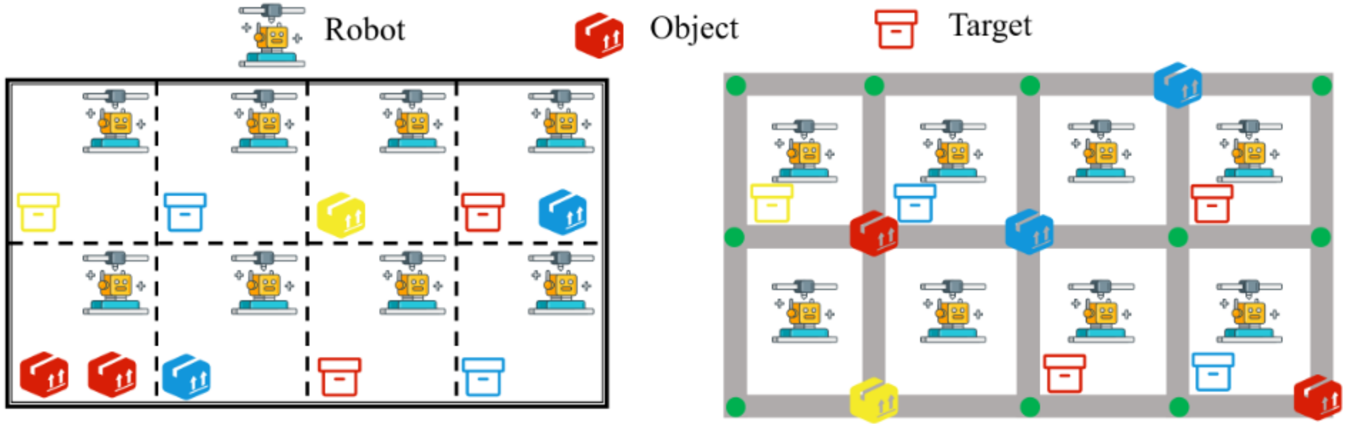


External Feedback

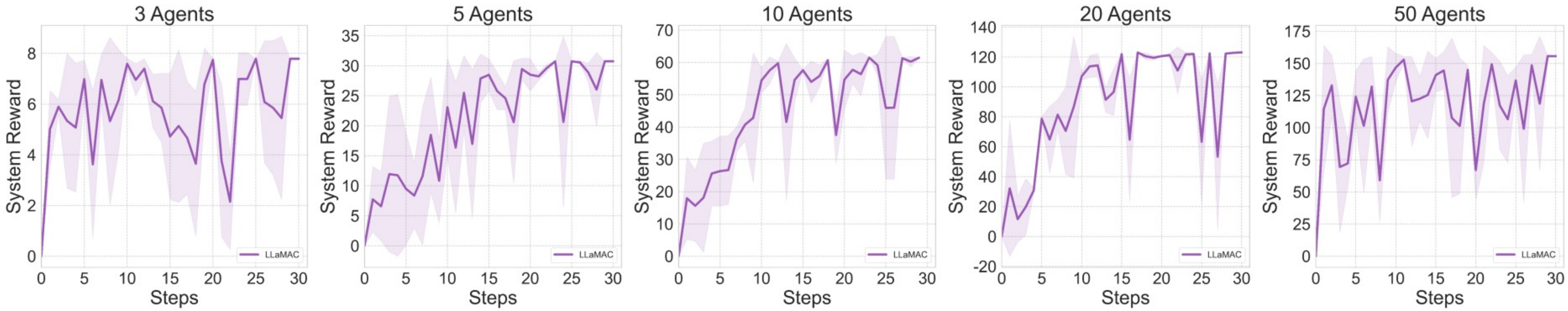
University of Chinese Academy of Sciences

Evaluation





System Resource Allocation $$G(x) = xe^{\frac{-(x-\mu)^2}{\sigma^2}}$$

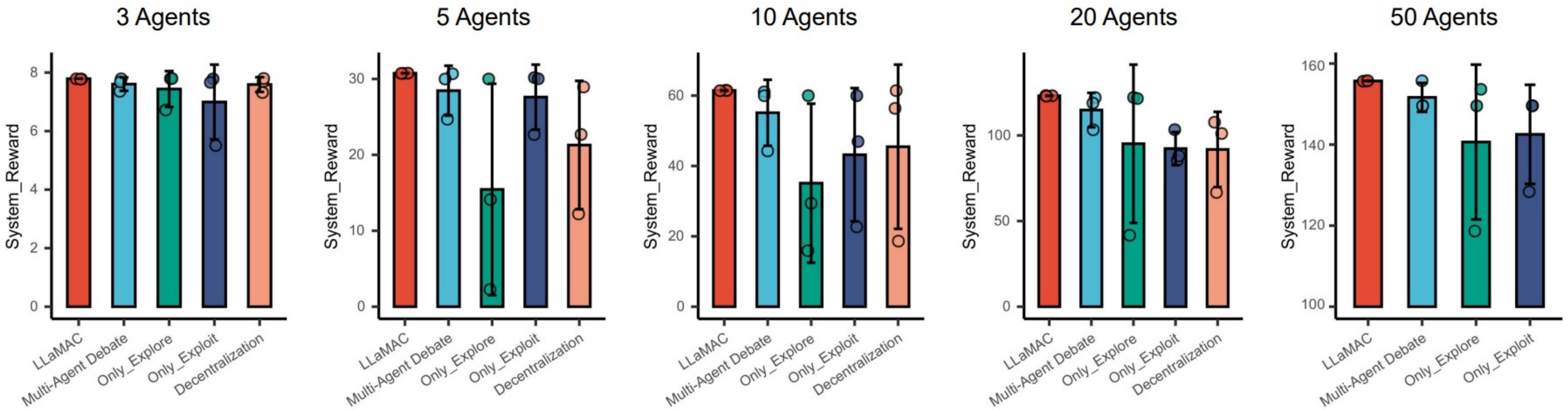


University of Chinese Academy of Sciences

Evaluation

System Resource Allocation

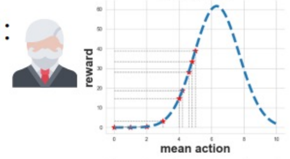
$$G(x) = xe^{\frac{-(x-\mu)^2}{\sigma^2}}$$



Step 0:

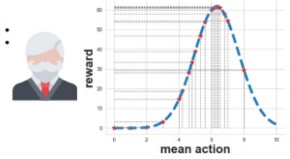
Given the limited data, **it's hard to infer a pattern or relationship** between action and system_reward. It's suggested to explore a higher mean_action.

Step 10:



The system reward seems to increase as the mean_action increases. The highest reward is achieved when the mean_action is 5.0. However, the rate of increase in reward seems to be slowing down as the mean_action increases, suggesting **a possible peak in the reward function.** To maximize rewards, it would be beneficial to explore slightly higher mean_actions to see if the reward continues to increase or starts to decrease.

Step 20:

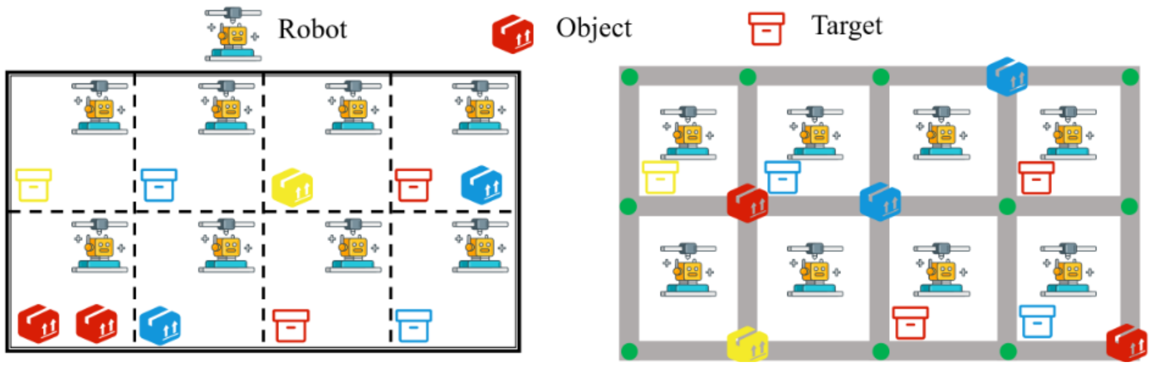


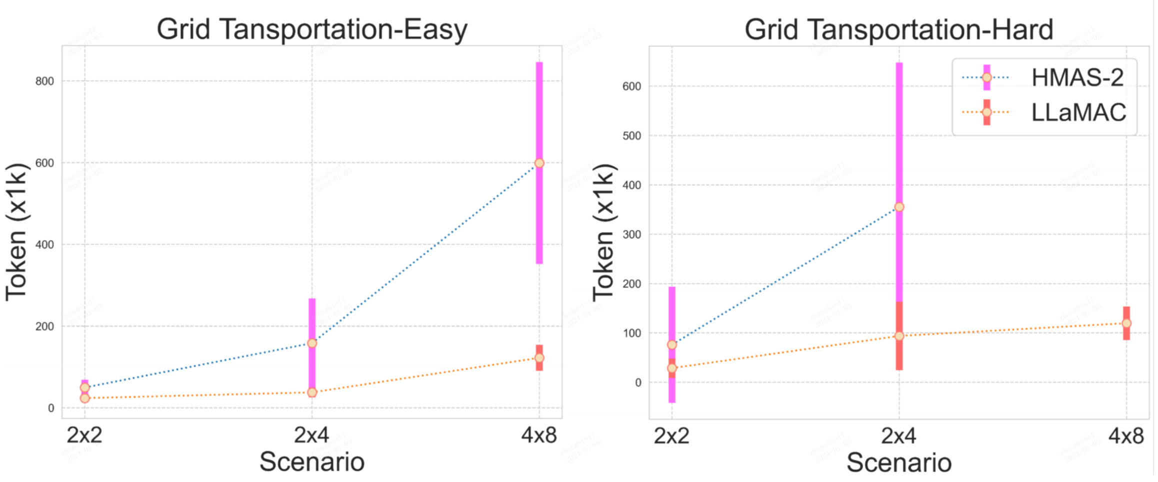
The system **reward seems to peak at an average action of 6.4, with a corresponding reward of approximately 61.49**. Both increases and decreases from this mean action value appear to result in lower rewards. Therefore, it seems that the optimal strategy is for all agents to choose actions that would result in an average value of around 6.4.

中国科学院大学
University of Chinese Academy of Sciences

Evaluation

Grid Transportation





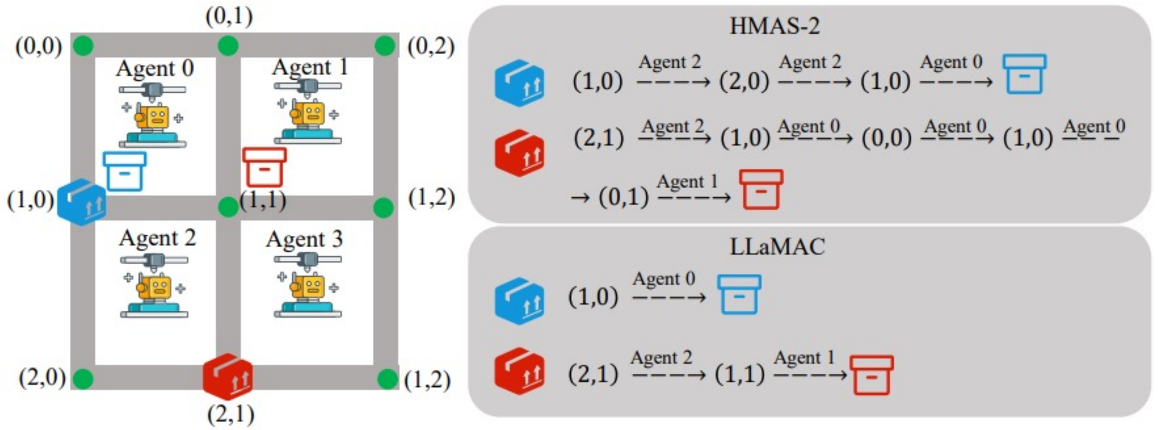


Table 2: Evaluation results under different grid settings in the Grid Transportation-Easy scene.

		Success	Steps	Feedback	Token($\times 1k$)
2x2	**HMAS-2**	100%	9.9(2.74)	3.3(2.05)	49.9(17.98)
	LLaMAC	**100%**	**7.0(1.79)**	**2.0(1.26)**	**23.9(8.38)**
2x4	**HMAS-2**	80%	15.5(6.09)	12.3(5.83)	158.4(107.84)
	LLaMAC	**100%**	**7.6(1.36)**	**4.3(1.42)**	**38.0(10.57)**
4x8	**HMAS-2**	60%	30.6(9.70)	26.1(13.59)	599.3(245.40)
	LLaMAC	**100%**	**12.9(2.70)**	**10.7(3.35)**	**122.6(30.55)**

Table 3: Evaluation results under different grid settings in the Grid Transportation-Hard scene.

		Success	Steps	Feedback	Token($\times 1k$)
2x2	**HMAS-2**	80%	7.0(5.0)	6.0(9.74)	76.1(116.66)
	LLaMAC	**100%**	**4.7(1.35)**	**3.6(2.80)**	**28.8(18.49)**
2x4	**HMAS-2**	20%	17.0(9.0)	24.0(20.0)	355.5(291.05)
	LLaMAC	**90%**	**7.44(2.95)**	**10.56(7.54)**	**94.0(68.09)**
4x8	**HMAS-2**	0%	-	-	-
	LLaMAC	**90%**	**8.44(1.57)**	**12.11(2.51)**	**119.8(32.75)**