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

2024.02.07
Lab Seminar
Park Kieun

HUMAN
CENTERED
COMPUTING
LABORATORY

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

Bin Zhang[1,2], **Hangyu Mao**[3,*], **Jingqing Ruan**[1,2], **Ying Wen**[4], **Yang Li**[5], **Shao Zhang**[4], **Zhiwei Xu**[1,2], **Dapeng Li**[1,2], **Ziyue Li**[3], **Rui Zhao**[3], **Lijuan Li**[1,2,*] and **Guoliang Fan**[1,2]

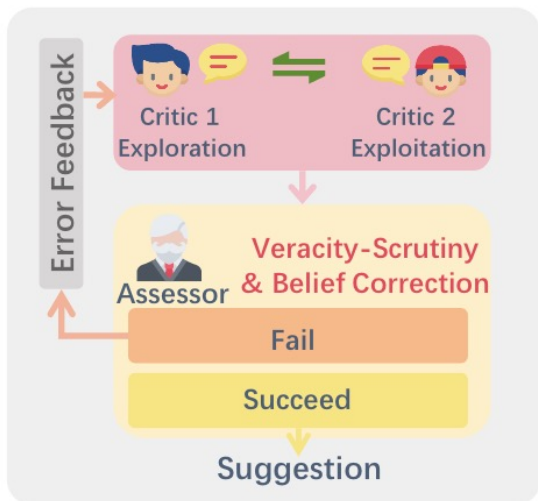
[1]Institute of Automation,Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

[3]SenseTime Research

[4]Shanghai Jiao Tong University

[5]The University of Manchester



HUMAN CENTERED COMPUTING LABORATORY

LLaMAC (Large **La**nguage **M**odel-based **A**ctor-**C**ritic)

- Task
  - Decision-making
    - System resource allocation
    - Grid transportation
- Method
  - MAS (multi-agent system) – max 50 agents experiment result
  - Actor-critic RL approach → TripletCritic structure
  - External feedback mechanism
  - Memory Module (=support Long-term memory)

LLaMAC framework

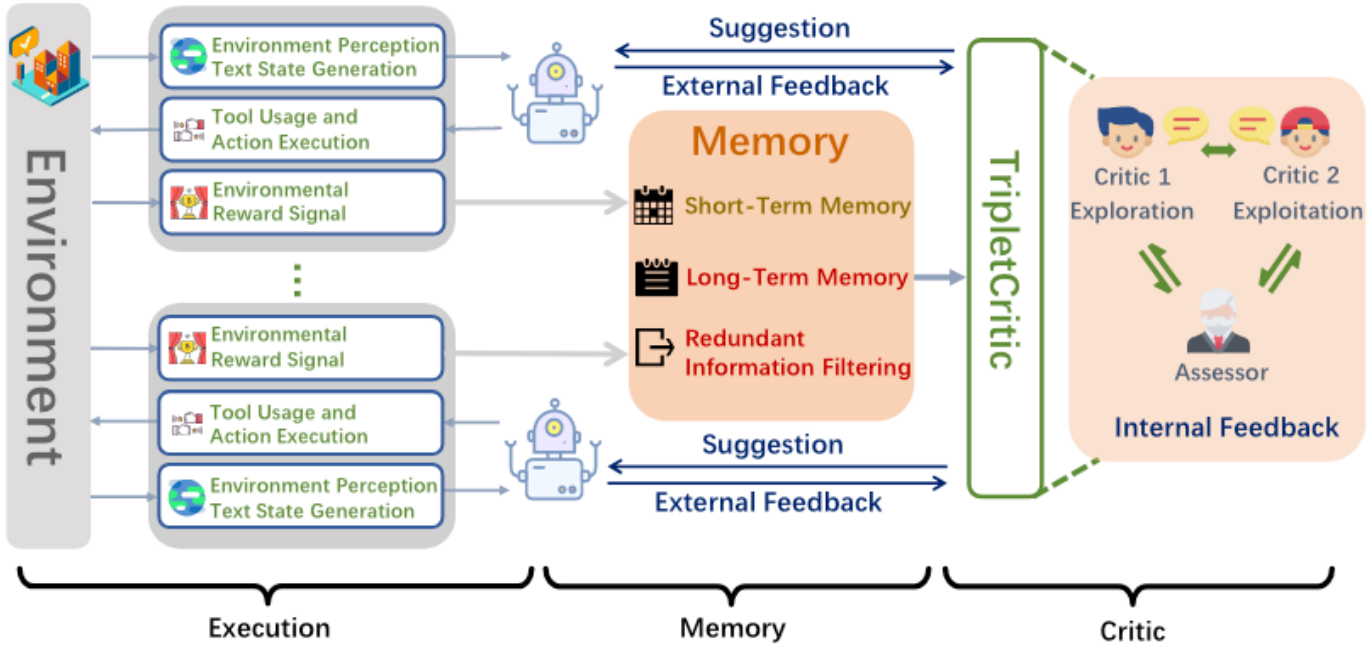


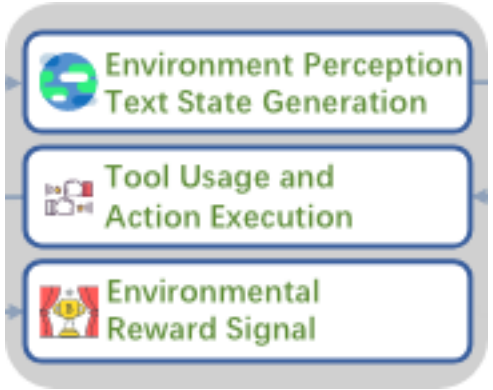
Figure 1: The overall framework of LLaMAC. The LLM-based agents achieve autonomous and continuous decision-making and interaction through the utilization of the execution, memory, and critic modules.

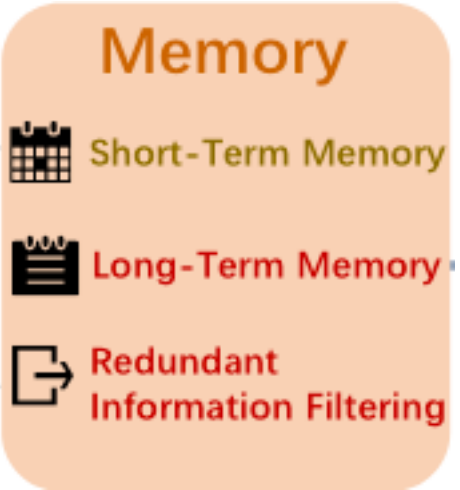
Execution & Memory module

- Execution module
  - Environment → text-based description

- Memory module
  - Short-term
    - Most recent state
  - Long-term
    - Most recent L steps of state transitions (state, action, reward)

$$< s_{t-L+1}, a_{t-L+1}, r_{t-L+1}, s_{t-L+2}, \ldots, s_t >$$





Critic module

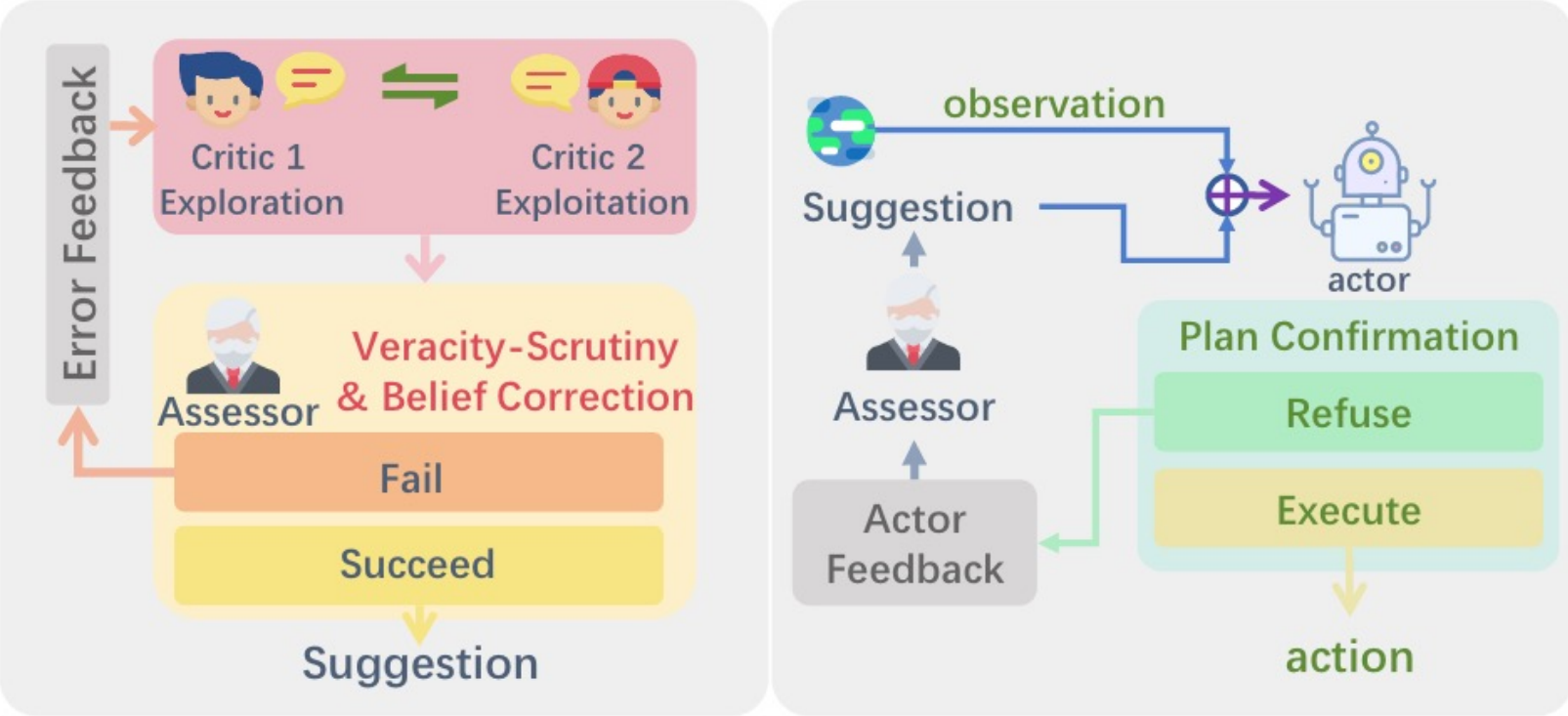
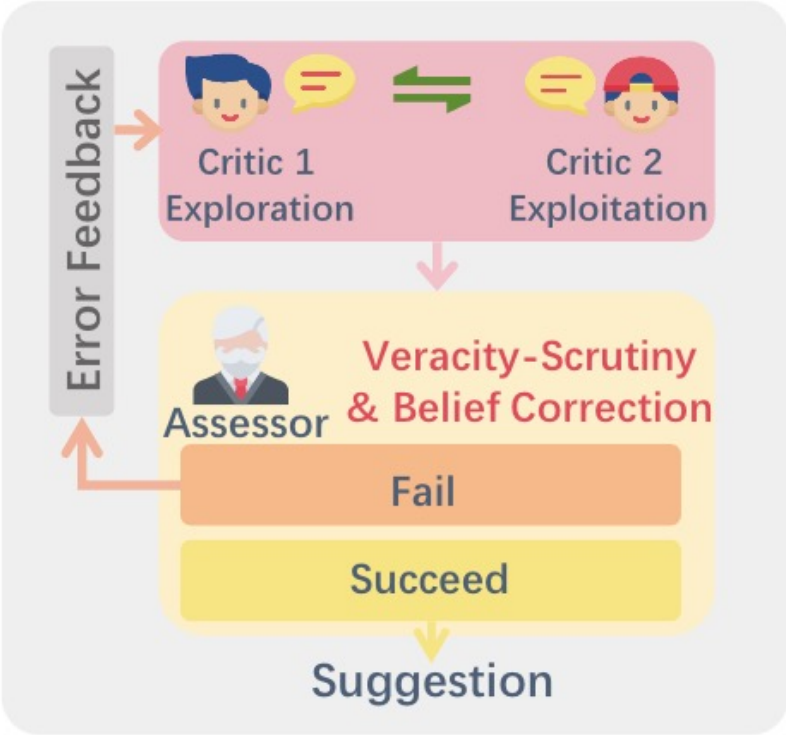


Figure 2: Internal Feedback within the TripletCritic (*Left*) and External Feedback mechanism from actor to critic (*Right*).

Critic module (TripletCritic : internal feedback)

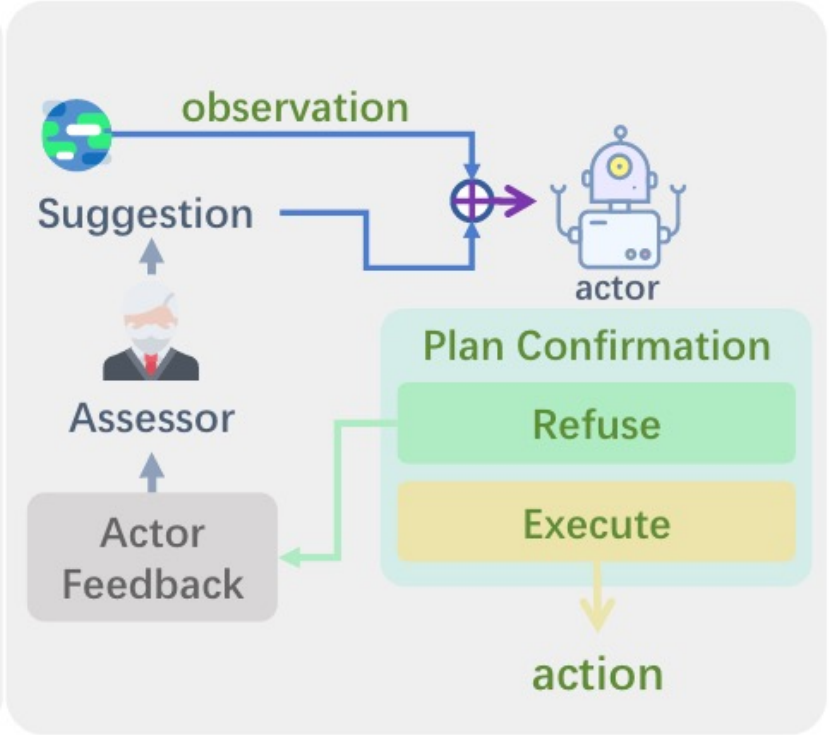
- Inspired by dopamine neurons in the brain [Dabney et al., 2020]

- Critic 1 : Exploration (장기적 목표에 중점)
- Critic 2 : Exploitation (단기적 목표에 중점)
- Assessor
  - Veracity-Scrutiny : 각 suggestions 에 오류가 없는지 확인
  - Belief Correction : 두 suggestions 의 균형점을 잡음
  - 최종 suggestion을 각 actor 에게 전달



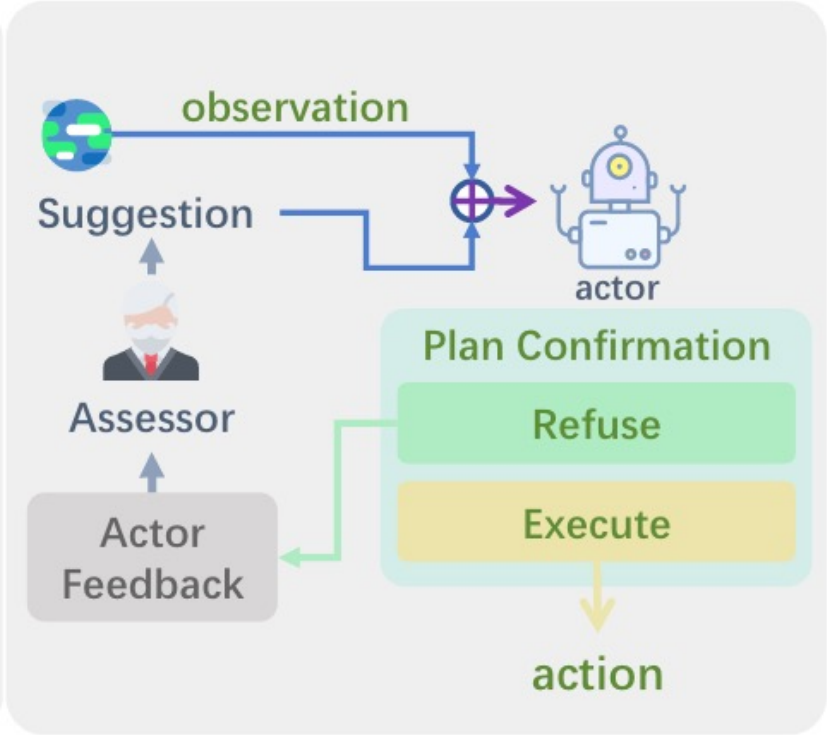
Critic module (external feedback flow)

1. TripletCritic : 각 actor 에게 suggestion 전달
2. Actor : observation 과 suggestion 을 바탕으로 실행 가능성 결정
3. 하나의 actor라도 반대한다면, 모든 actor 의 feedback을 모아 TripletCritic 에 feedback
4. 모든 actor 가 suggestion 에 만족할 때까지 반복



Critic module (external feedback advantage)

- Internal + External feedback
  - Form a comprehensive & automated iterative process
- Reduce hallucination
- Reliability of TripletCritic reduce external feedback of actors
  - Minimize access cost
  - Token-efficient



LLaMAC flow

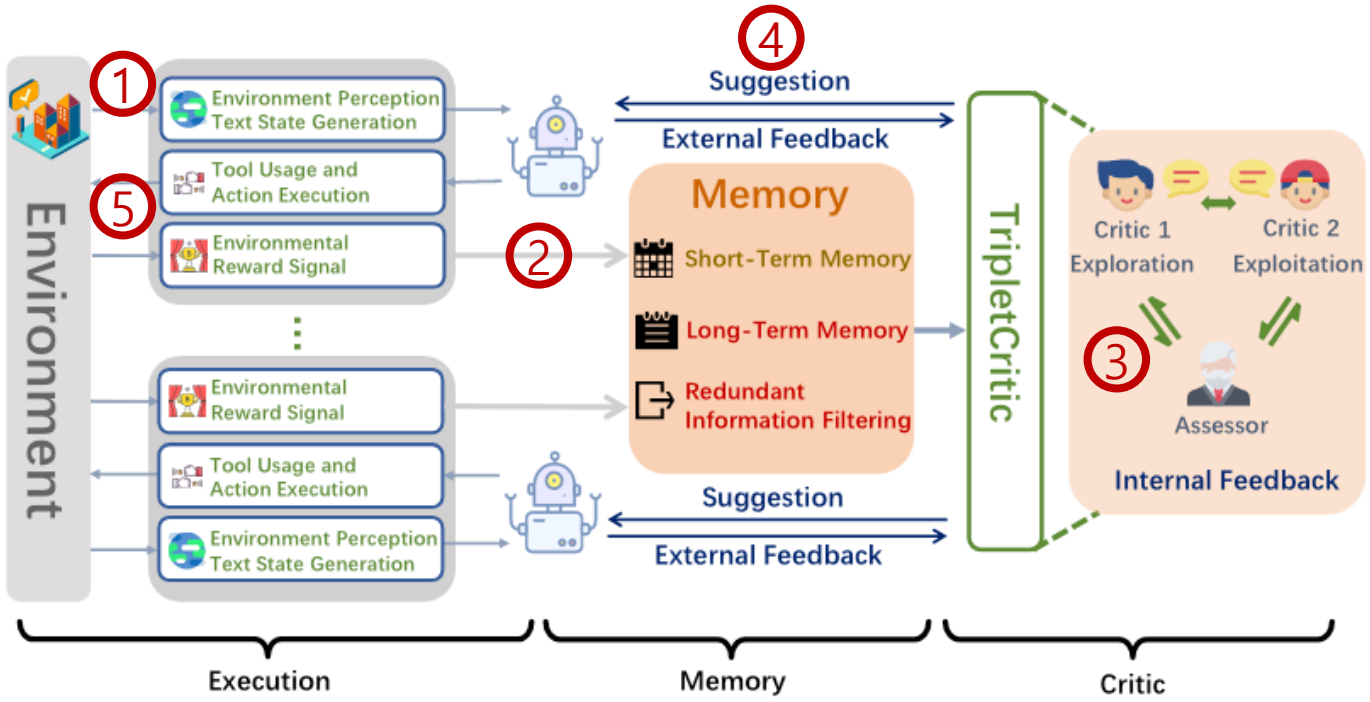
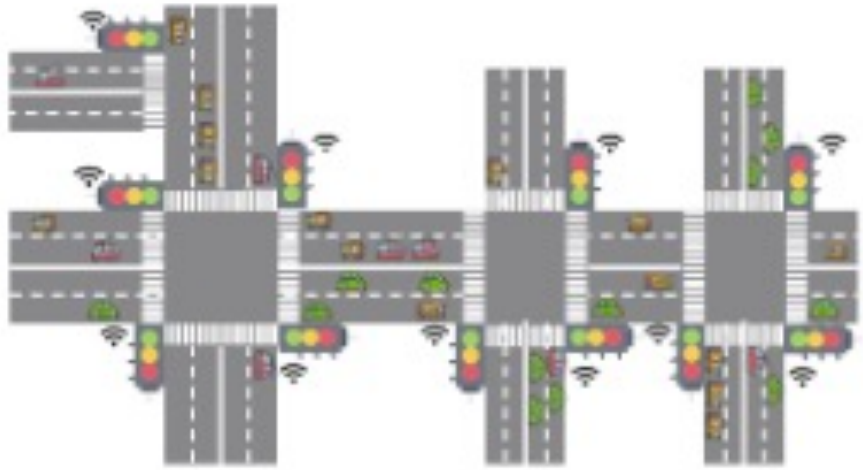


Figure 1: The overall framework of LLaMAC. The LLM-based agents achieve autonomous and continuous decision-making and interaction through the utilization of the execution, memory, and critic modules.

System resource allocation

- Traffic control
  - Single-step decision
  - Require mathematical reasoning capabilities

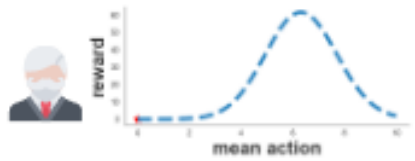


- Multiple traffic controllers (agents)
  - Select an integer between 0~9
  - No knowledge of the choices made by other agent

System resource allocation

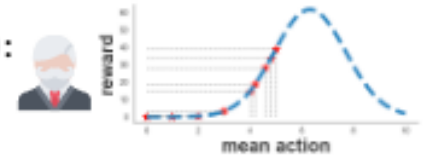
- Cognitive process of the assessor

Step 0:



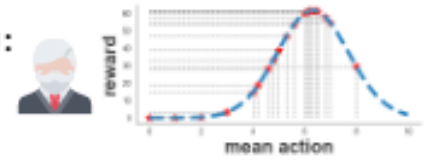
Given the limited data, **it's hard to infer a pattern or relationship** between action and system_reward. It's suggested to explore a higher mean_action.

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.

System resource allocation

- LLaMAC - stable performance
- Multi-agent Debate - converge to local optima

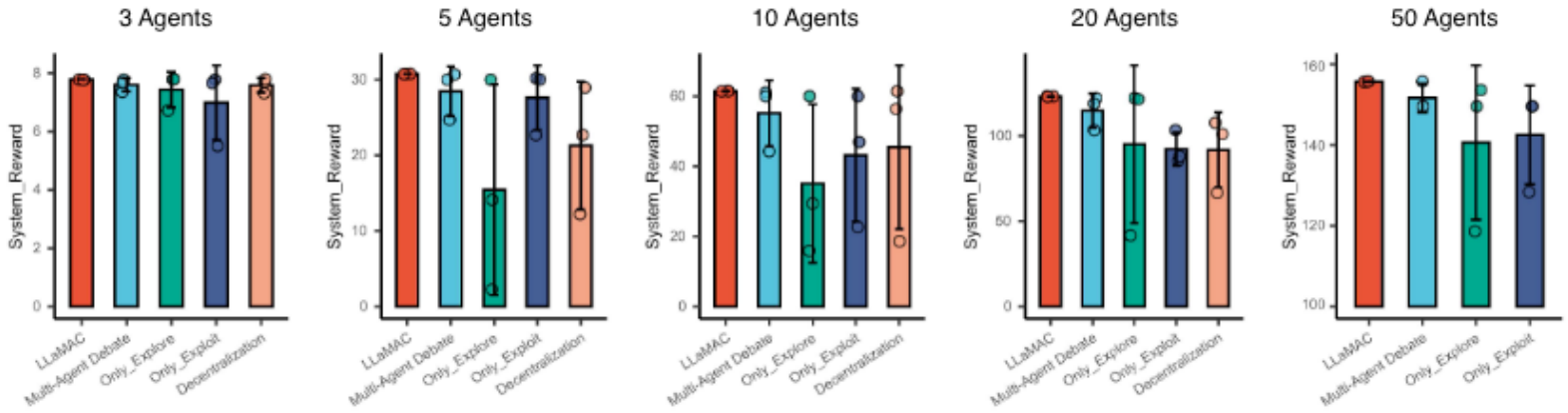


Figure 5: The final performance of different methods in system resource allocation scenarios with different number of agents.

Grid transformation

- Robot grid transformation
  - Multi-step decision
  - Require spatial & logical reasoning
  - Requrie long-term planning ability
- Mode
  - Easy : moving across the grid
  - Hard
    - moving along the grid boundaries
    - Possibility of Conflict & Collision

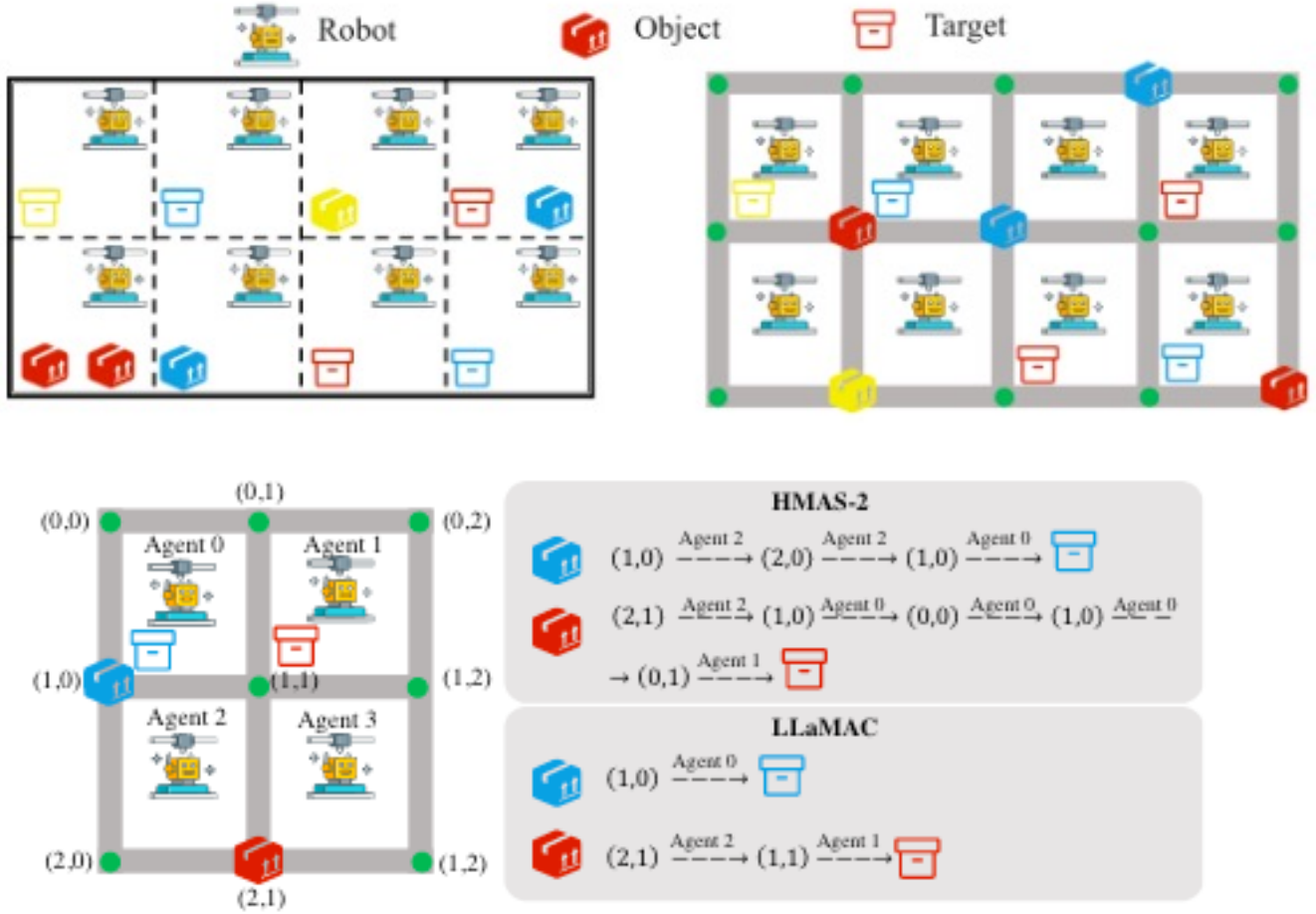


Figure 8: The performance of LLaMAC and HMAS-2 in the 2x2 robotic grid transportation scenario. To enhance visualization, non-essential objects and targets within the scene are concealed.

Grid transformation

- SOTA
  - long-term planning & execution
  - Spatial reasoning
  - Learning from interactions or errors

Table 2: Evaluation results under different grid settings in the Grid Transportation scenarios include metrics such as the success rate (*Success*), time steps (*Steps*) taken to execute tasks, and the count of feedback instances (*Feedback*). The values in parentheses correspond to a single standard deviation over 10 trials.

		Grid Transportation-Easy			Grid Transportation-Hard		
		Success	**Steps**	**Feedback**	**Success**	**Steps**	**Feedback**
2×2	**HMAS-2**	**100%**	9.9(2.74)	3.3(2.05)	80%	7.0(5.0)	6.0(9.74)
	LLaMAC	**100%**	**7.0(1.79)**	**2.0(1.26)**	**100%**	**4.7(1.35)**	**3.6(2.80)**
2×4	**HMAS-2**	80%	15.5(6.09)	12.3(5.83)	20%	17.0(9.0)	24.0(20.0)
	LLaMAC	**100%**	**7.6(1.36)**	**4.3(1.42)**	**90%**	**7.44(2.95)**	**10.56(7.54)**
4×8	**HMAS-2**	60%	30.6(9.70)	26.1(13.59)	0%	-	-
	LLaMAC	**100%**	**12.9(2.70)**	**10.7(3.35)**	**90%**	**8.44(1.57)**	**12.11(2.51)**

Contributions

- Introduce LLaMAC
  - Internal feedback mechanism (TripletCritic)
  - External feedback mechanism
- Remarkable performance (SOTA)
- Reduce access cost

QA

HUMAN
CENTERED
COMPUTING
LABORATORY