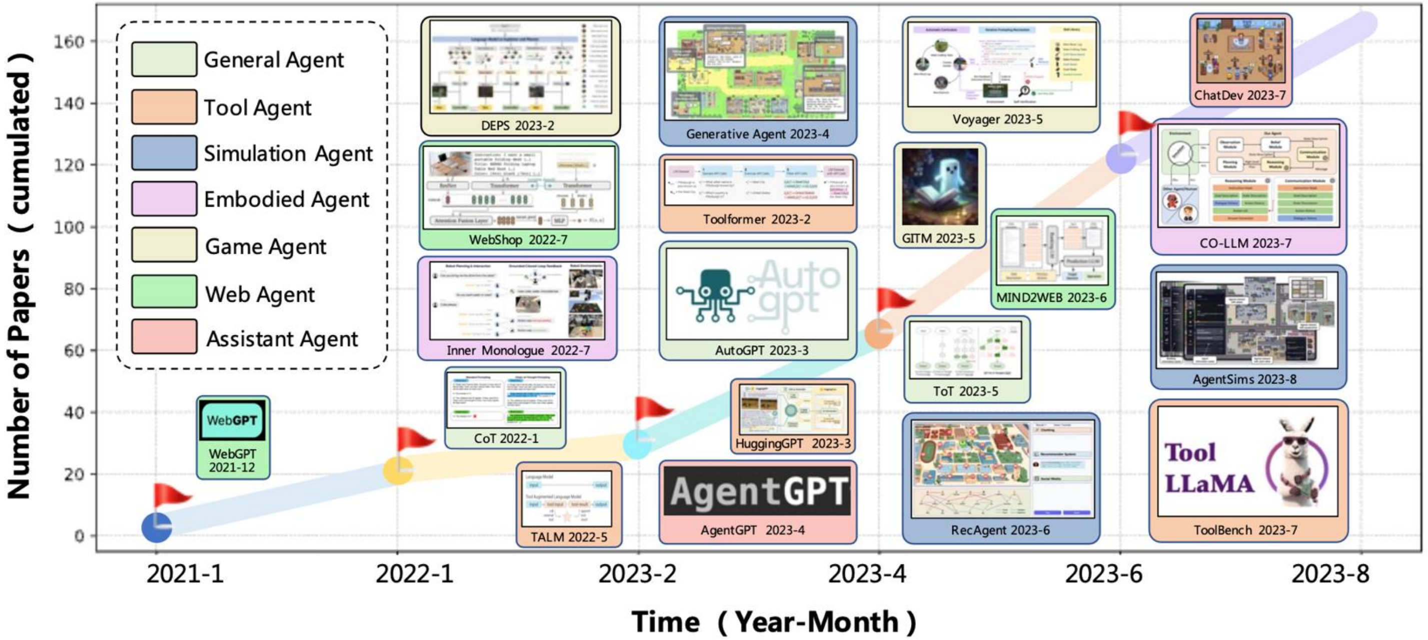
中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Reflective Multi-Agent Collaboration based on Large Language Models

Reflective Multi-Agent Collaboration based on Large Language Models

- Benefiting from the powerful language expression and planning capabilities of Large Language Models, LLM-based autonomous agents have achieved promising performance in various downstream tasks
- Recently, based on the development of single-agent systems, researchers propose to construct multi-agent systems in response to the growing task complexity



Reflective Multi-Agent Collaboration based on Large Language Models

In this paper, we consider leveraging the self-reflection mechanism to improve multi-agent collaboration, and propose an elegant framework named COPPER.

- **Agent Decision Process**

$$a_{k,t} = \text{Actor}^i(p^i, sm^i_{k,t}, s_{k,t}).$$

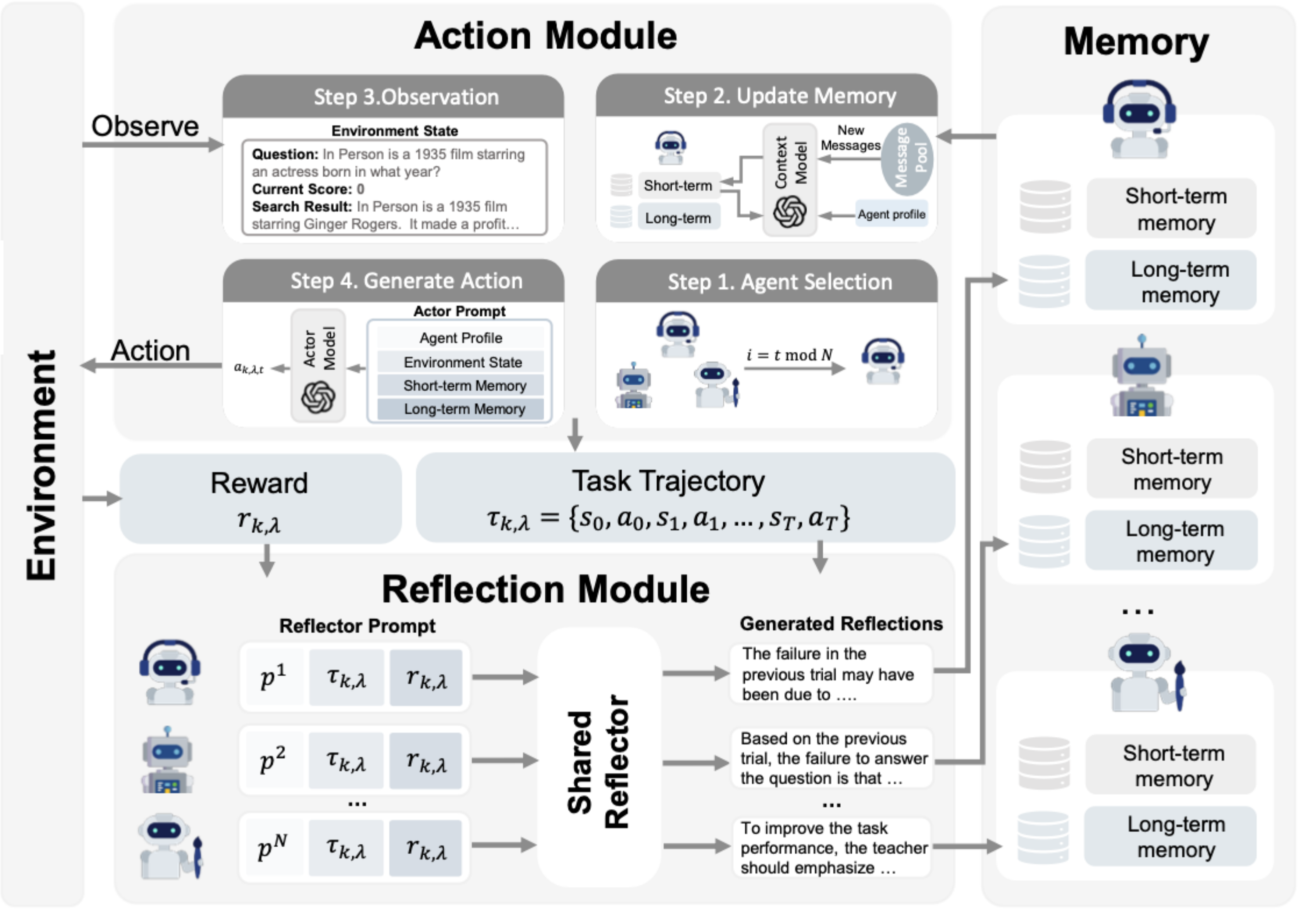
$$a_{k,\lambda,t} = \text{Actor}^i(p^i, lm^i_{k,\lambda}, sm^i_{k,\lambda,t}, s_{k,\lambda,t}),$$

- **Agent Reflection Process**

$$y^i_{k,\lambda} = \text{Reflector}^i(p^i, [sm^i_{k,\lambda,T}]^N_{i=1}, r_{k,\lambda}),$$

- **Update of Short-Term Memory**

$$sm^i_{k,t} = \text{Context}^i(p^i, sm^i_{k,t-1}, \{s_i, a_i\}^t_{i=\max(0,t-N+1)}),$$



高瓴人工智能学院
Gaoling School of Artificial Intelligence

Reflective Multi-Agent Collaboration based on Large Language Models

- Towards more efficient reflection, we propose to train a shared reflector using the counterfactual Proximal Policy Optimization (PPO) mechanism.
  - The counterfactual reward can be evaluated according to the impact of each agent reflection on enhancing the overall task performance.
  - To enhance the training efficiency and stability, we gather reflection data across agents and propose to train a shared reflector.

Reflective Multi-Agent Collaboration based on Large Language Models

- **Construction of Counterfactual Reward**

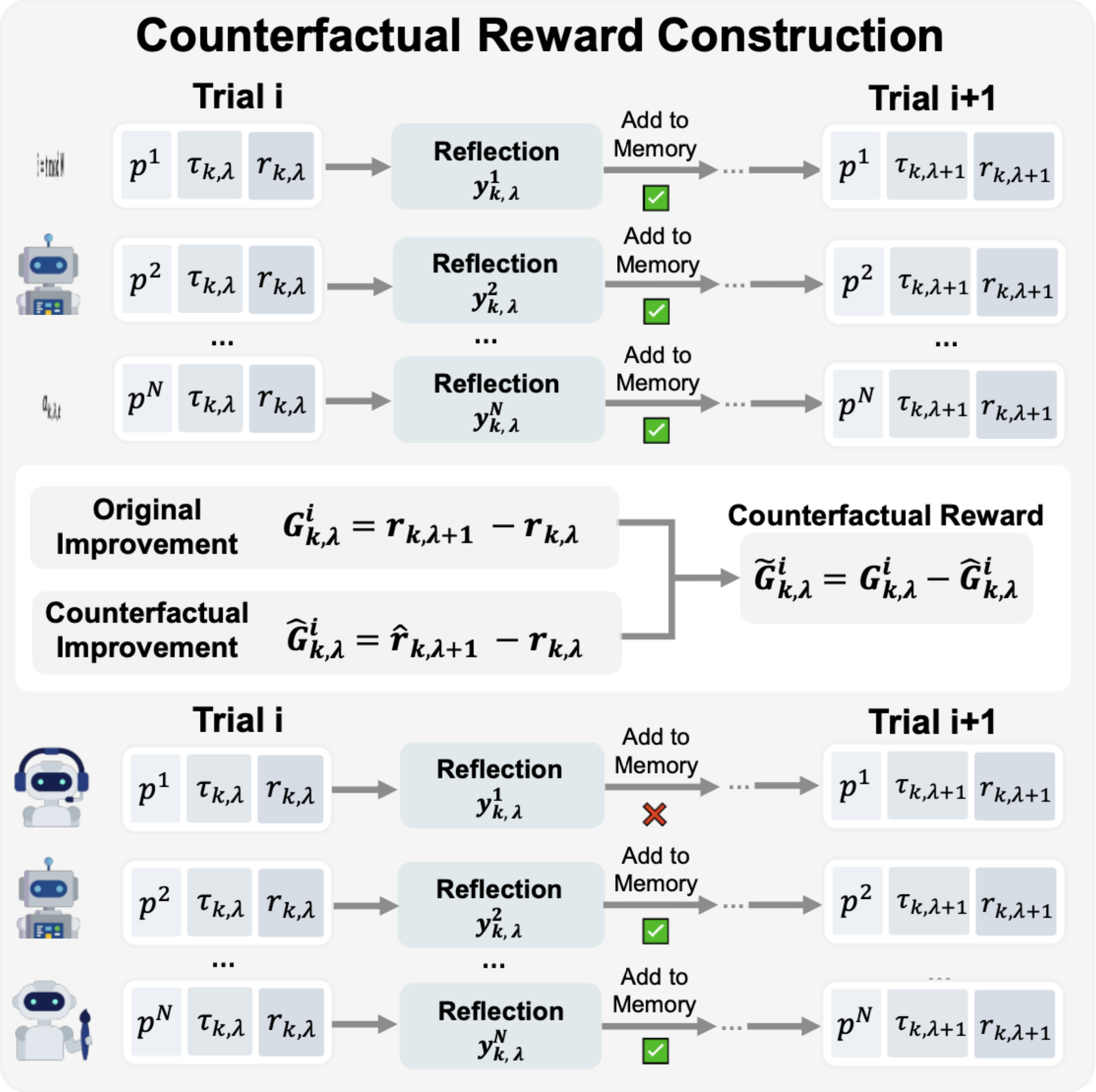
$$G^i_{k,\lambda} = r_{k,\lambda+1} - r_{k,\lambda}$$

$$\hat{G}^i_{k,\lambda} = \hat{r}_{k,\lambda+1} - r_{k,\lambda}$$

$$\tilde{G}^i_{k,\lambda} = G^i_{k,\lambda} - \hat{G}^i_{k,\lambda}.$$

- **Training Data Collection**

$$D_{CF} = \{(x^i_{k,\lambda}, y^i_{k,\lambda}, \tilde{G}^i_{k,\lambda}) \\ | 1 \le i \le N, 1 \le \lambda \le \Lambda, 1 \le k \le K\}$$



Reflective Multi-Agent Collaboration based on Large Language Models

- We follow the Reinforcement Learning from Human Feedback (RLHF) method and adopt a similar three-step approach to fine-tune the shared reflector with counterfactual rewards.

  - **Supervised Fine-Tuning**

$$\mathcal{L}_{SFT}(\boldsymbol{\theta}) = -\mathbb{E}_{(x,y)\sim D_{CF}}[\sum_{k=1}^{m} \log \pi_\theta(y_k|x, y_{<k})],$$

  - **Training Reward Model**

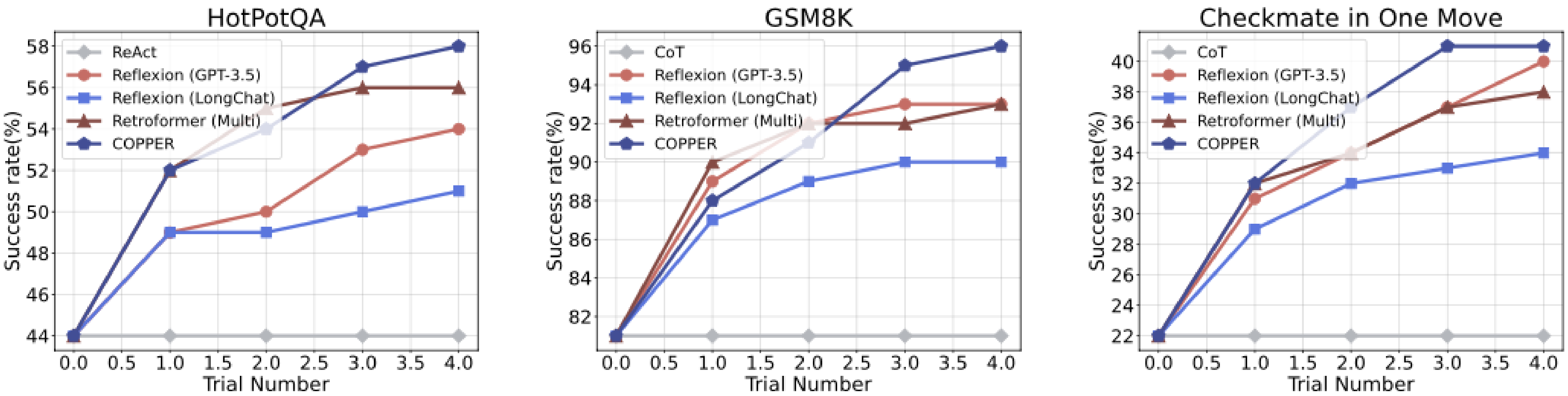
$$\mathcal{L}_{RM}(\boldsymbol{\phi}) = \mathbb{E}_{(x,y,r)\sim D_{CF}}[(R_{CF_\phi}(x,y) - r)^2].$$

  - **Proximal Policy Optimization**

$$\mathcal{L}_{PPO}(\boldsymbol{\theta}) = -\mathbb{E}_{x\sim D_{CF}}\mathbb{E}_{y\sim\pi_\theta^{RL}(x)}[R_{CF_\phi}(x,y) - \beta \log \frac{\pi_\theta^{RL}(y|x)}{\pi^{SFT}(y|x)}].$$

Reflective Multi-Agent Collaboration based on Large Language Models

We compare the performance of COPPER against different baselines on HotPotQA, GSM8K, and Checkmate in One Move datasets. Experimental results indicate that our COPPER exhibits superior reflective ability in multi-agent collaboration scenarios.



(a) HotpotQA. (b) GSM8K. (c) Checkmate in One Move.

Figure 3: Performance of COPPER against baselines on three datasets.

THANKS!