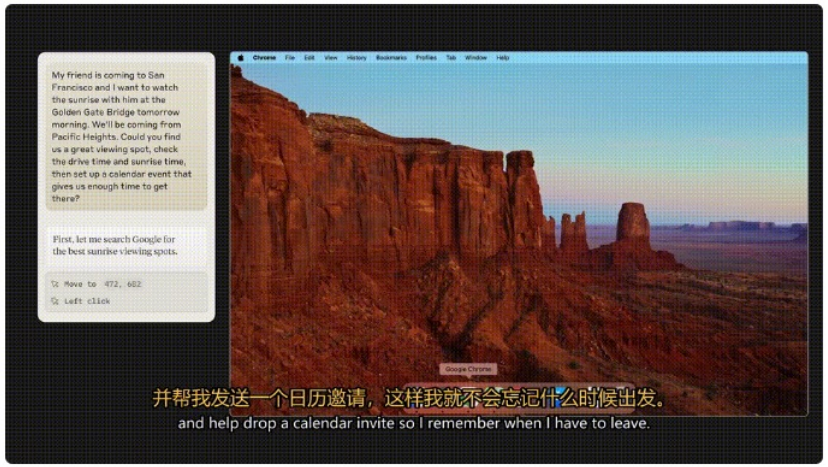
面向可交互环境泛化的多智能体强化学习框架

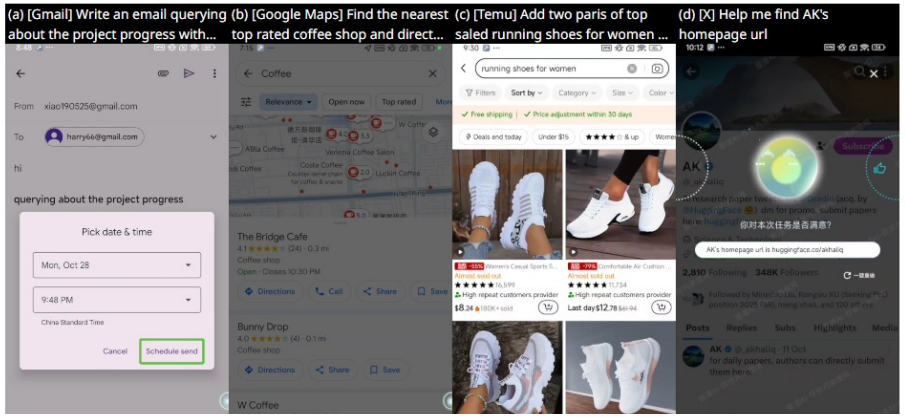
刘子君

清华大学计算机系
2025.4.7

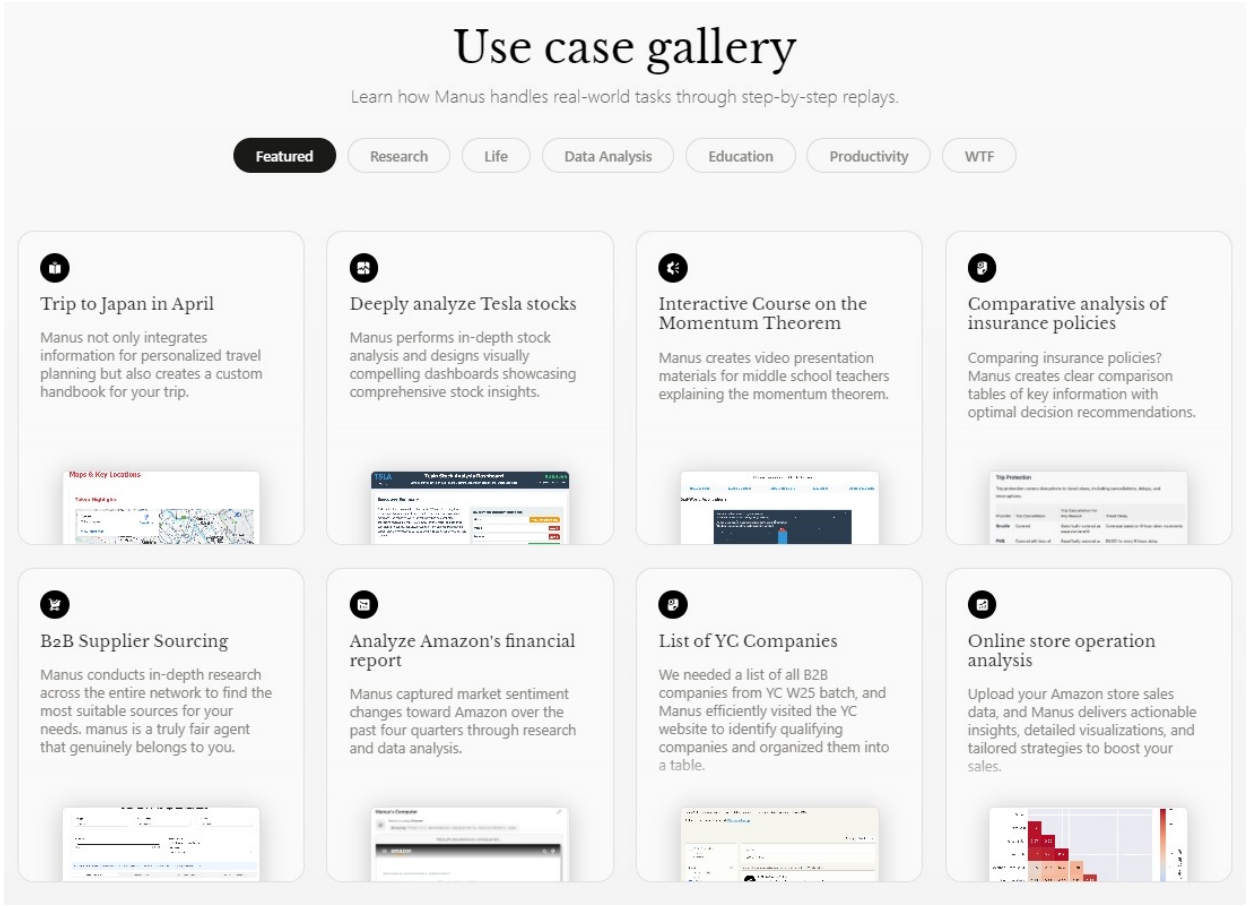
背景：GUI环境交互智能体逐步落地



Claude “Computer Use”



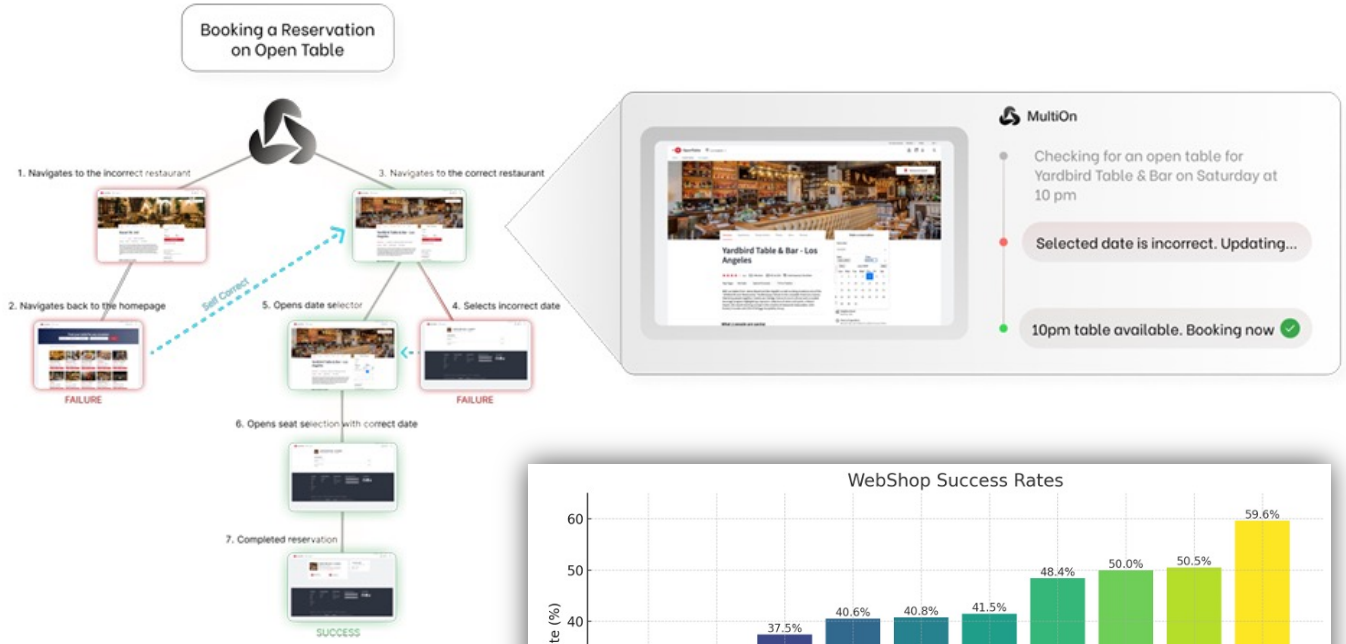
AutoGLM “Mobile Use”

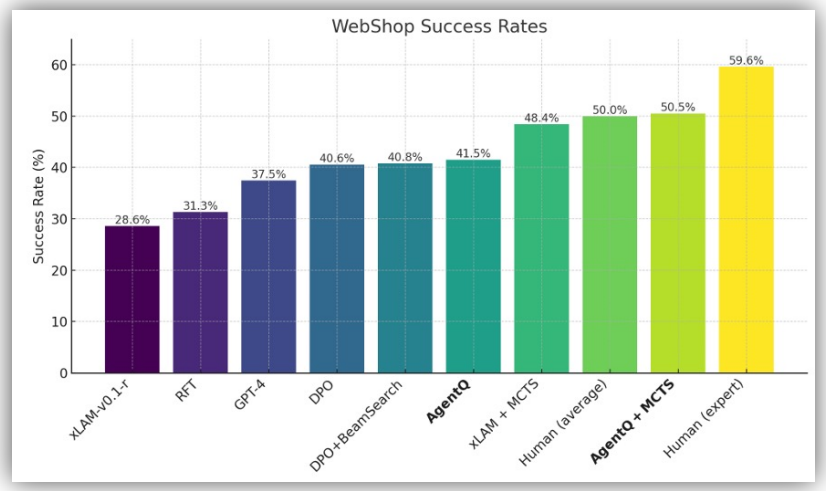


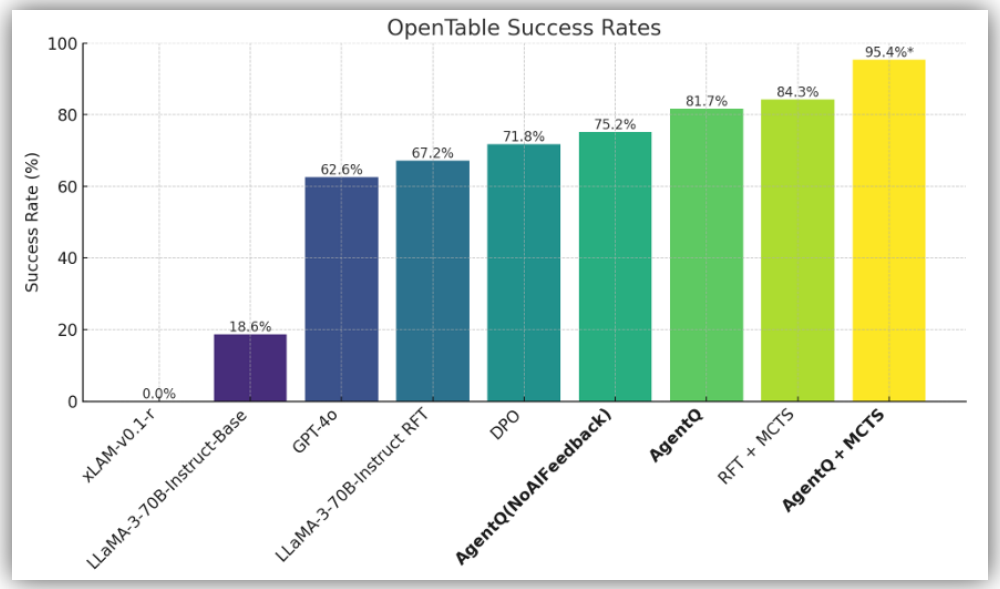
Manus “General UI Manipulation”

Liu et al. AutoGLM: Autonomous Foundation Agents for GUIs. ArXiv e-print. 2024.

背景：面向可交互环境的单智能体训练方法泛化性差





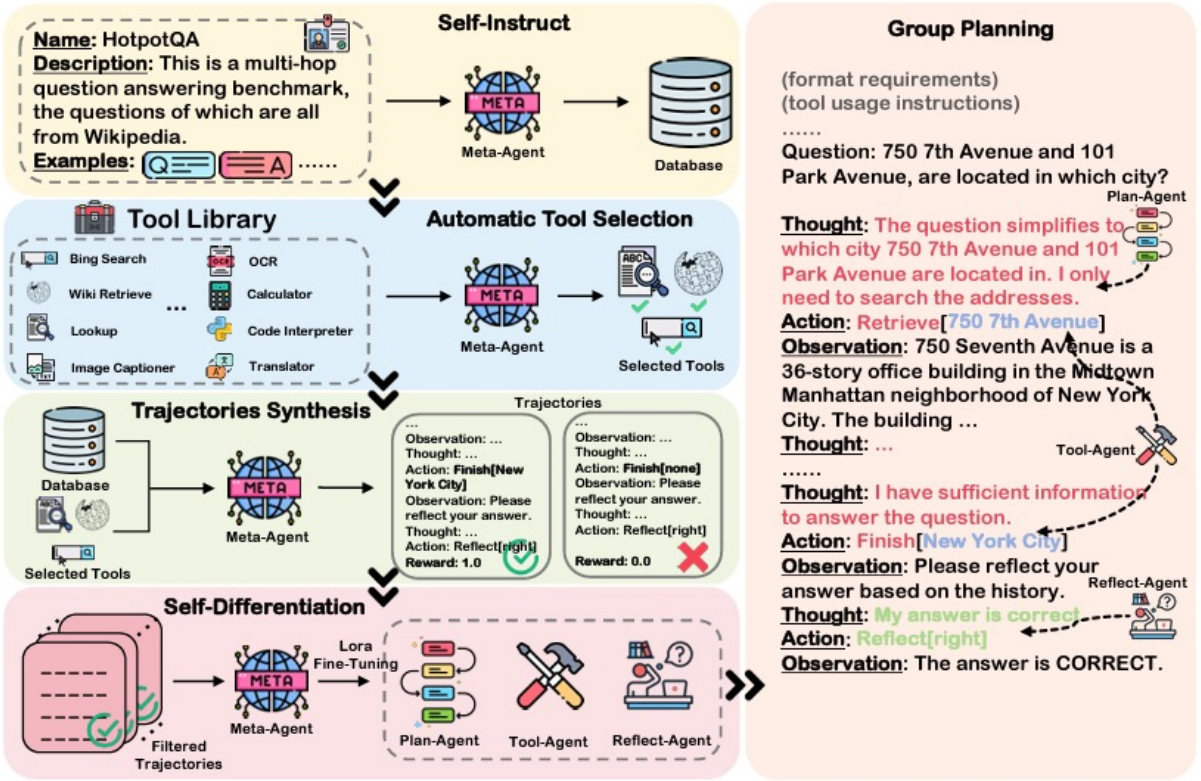


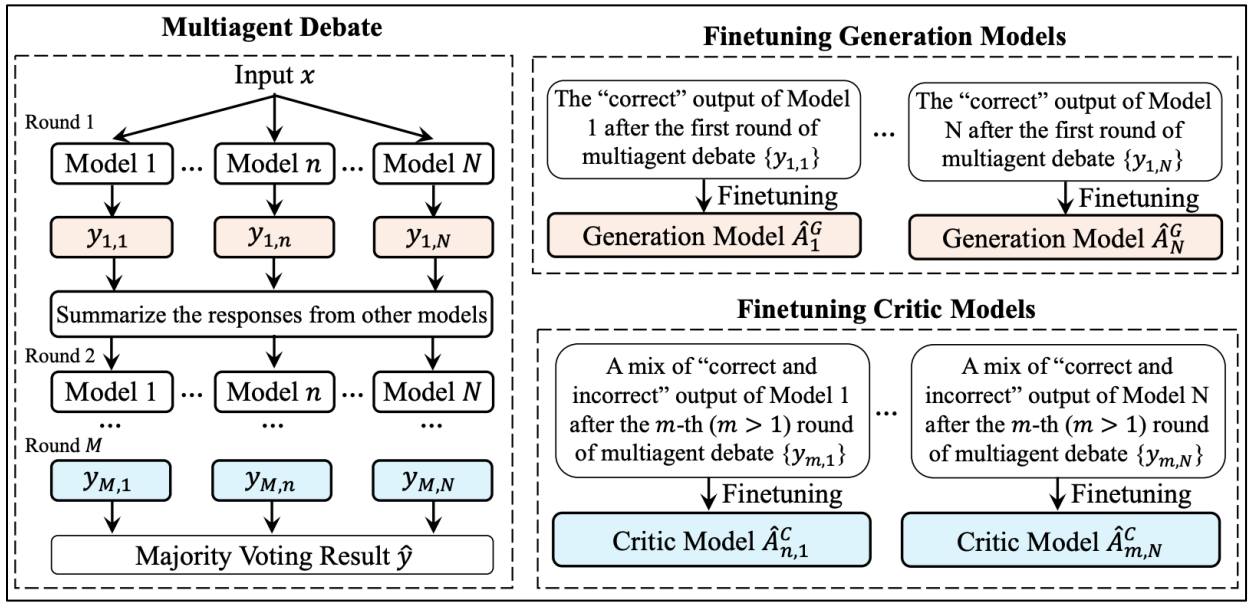
通过在可交互环境上训练单智能体（RFT、DPO、PG...），模型的智能体能力有所提升

当前单智能体训练方法在环境间**泛化性差**

Putta et al. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents. ArXiv e-print. 2024.

背景：现有多智能体训练方法依赖人类先验



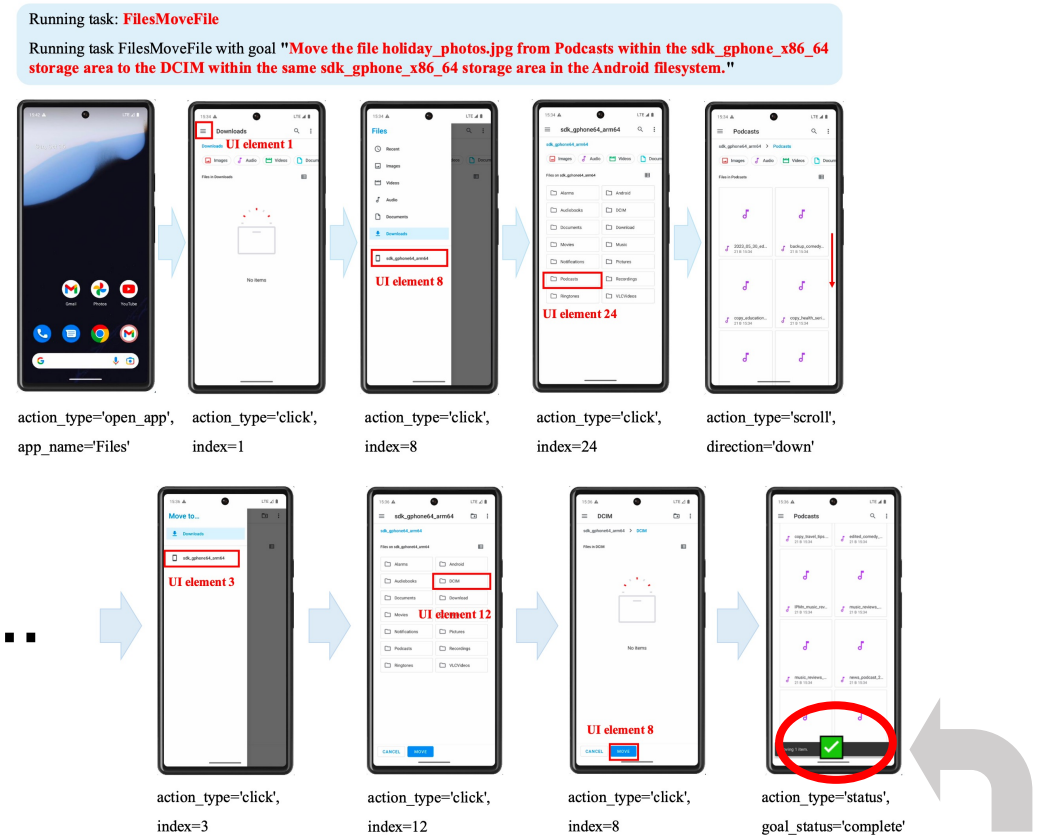


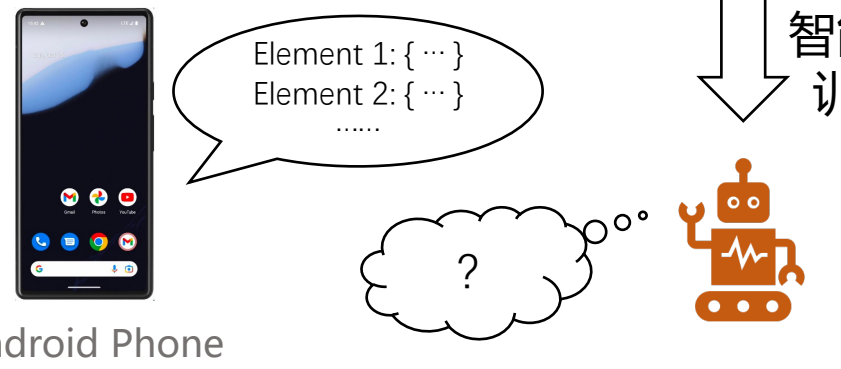
人工分配智能体角色，进行微调（RFT），适配单一任务或多智能体系统架构

Qiao et al. AutoAct: Automatic Agent Learning from Scratch for QA via Self-Planning. ACL 2024. Subramaniam et al. Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains. ICLR 2025.

技术挑战：智能体训练方法对于可交互环境适配性低

- 智能体模型训练仅依赖于到端的奖励且信号稀疏，导致模型性能与泛化较差。





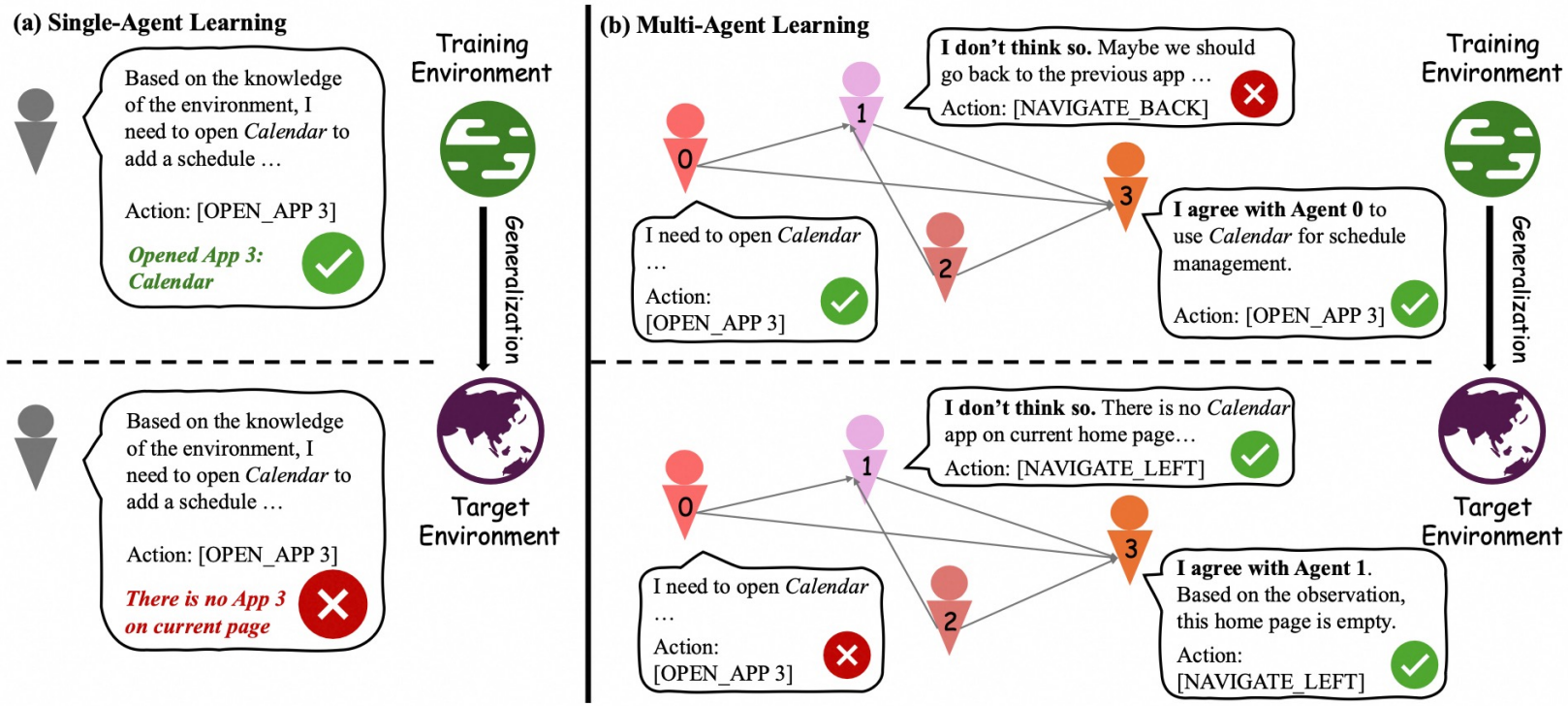
智能体训练

多轮交互中，**仅最终交互结果具有奖励反馈信号**，缺乏过程性监督，训练所得智能体性能较差

到端奖励只评价智能体在训练环境内的表现，智能体易**过度学习训练环境专属的知识**，导致泛化性差

启发：智能体合作的潜在泛化性

- 部分环境专属的知识难以泛化，但智能体之间互相合作的方式多数可以泛化。



环境内容物等知识难以泛化，但智能体之间纠正、互助、引申等合作形式可以泛化

He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

启发：智能体合作的潜在泛化性

- 部分环境专属的知识难以泛化，但智能体之间互相合作的方式多数可以泛化。

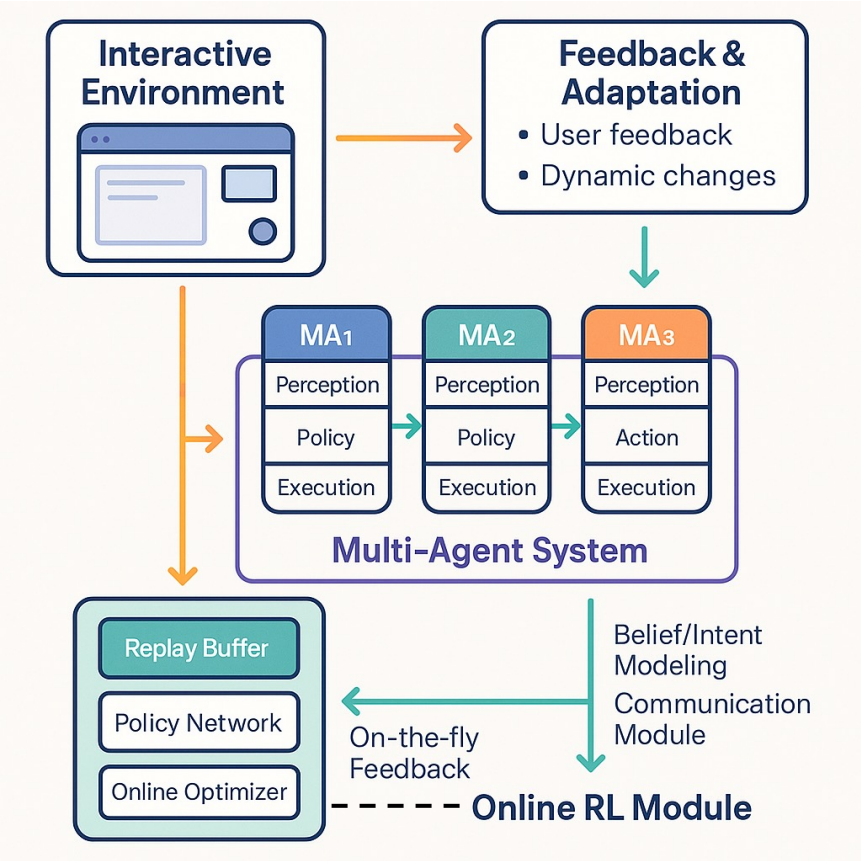
如何同时实现训练环境上的性能提升与环境间的高效泛化？

环境内容物等知识难以泛化，但智能体之间纠正、互助、引申等合作形式可以泛化

He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

基本框架：面向可交互环境的多智能体在线强化学习

- **不指定智能体角色，在线探索环境，基于奖励优化更新模型参数、系统结构**

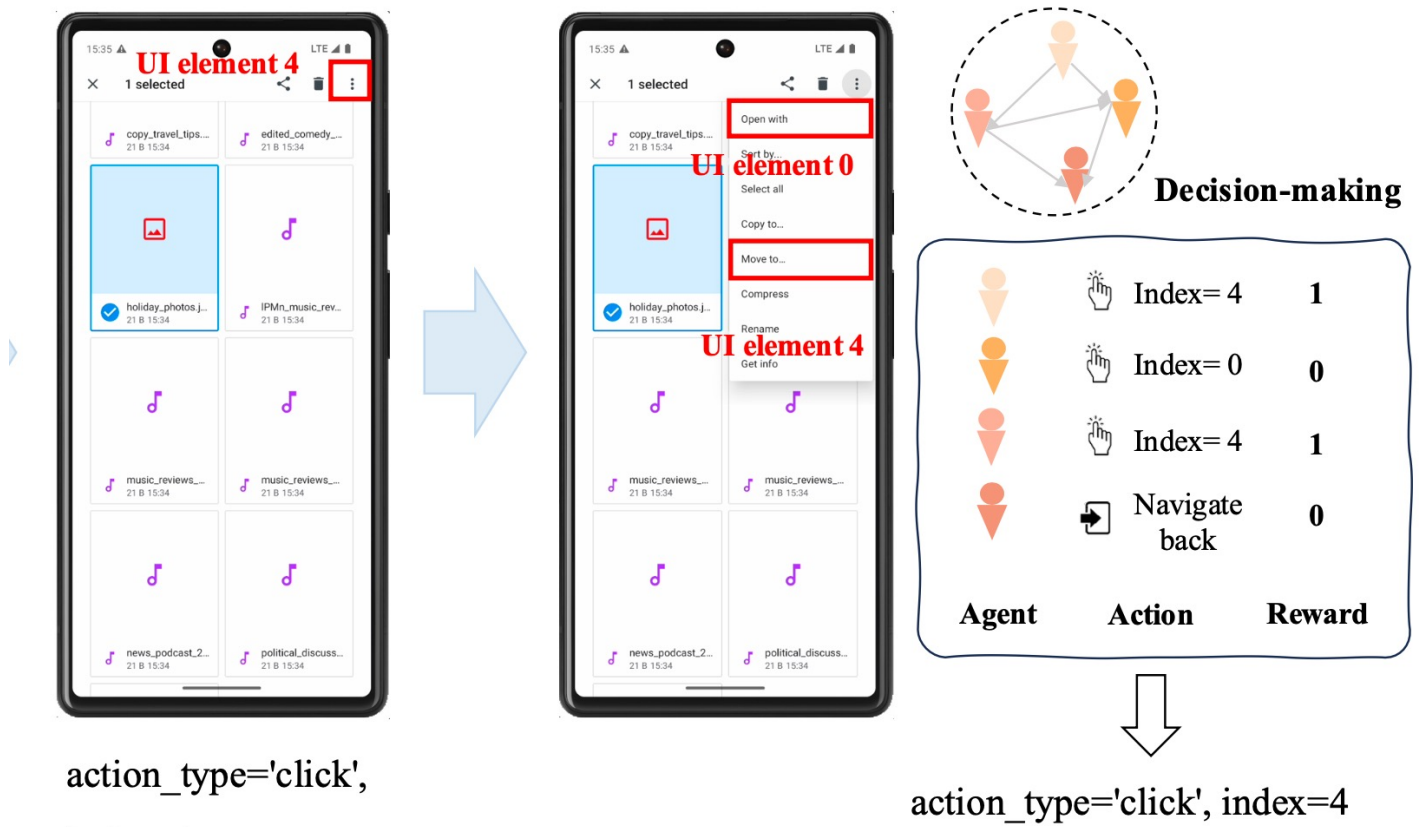


与传统MARL方法的区别：

- Token级价值网络（Value Network）的训练和更新开销过大，且优化较不稳定；现有工作往往通过样本级价值网络或DPO类方法**避免引入该部分训练。**
- 传统多智能体强化学习较少考虑智能体间通信，以及如何在线更新系统架构拓扑，**限制了语言多智能体系统的强化学习框架的完整性。**
- 传统多智能体强化学习定义的动作空间更清晰，语言多智能体同时生成交流与决策内容，需要设计方法**同时考虑交流与决策的贡献，并进行同步优化。**

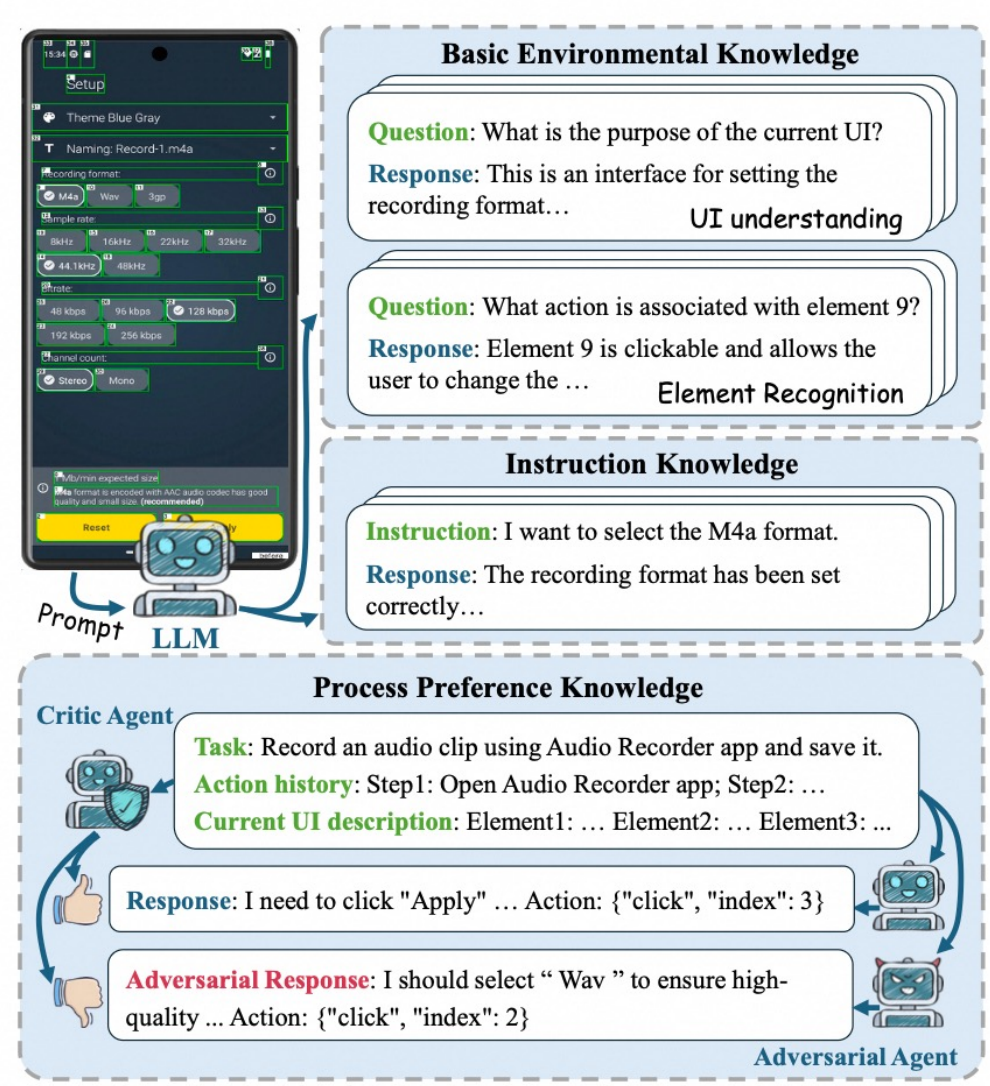
创新思路：基于多智能体引导信用重分配策略的强化学习

- **引入训练时多智能体，基于通用知识重分配奖励信号，并对抗构造偏好样本**



- 对于多智能体策略的每次响应，评判智能体基于大模型预训练知识重新给出**细化的过程奖励信号**，从而替代原本到端奖励信号
- 对于每个智能体策略网络做出的决策，首先标注二元过程奖励标签，对于正决策样本，对抗智能体基于拒绝采样**提出对抗性副样本，进而实现偏好优化**

技术方案：通用环境知识训练（阶段一）



- **对通用环境知识进行指令数据合成，进而实现多阶段单智能体课程学习**

指令微调阶段

1. 合成基础环境信息相关的问答数据进行训练
2. 合成简单用户指令数据进行训练

偏好学习阶段

1. 在环境中采集成功轨迹样本
2. 利用对抗智能体，合成负面轨迹样本

He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

技术方案：多智能体在线强化学习（阶段二）

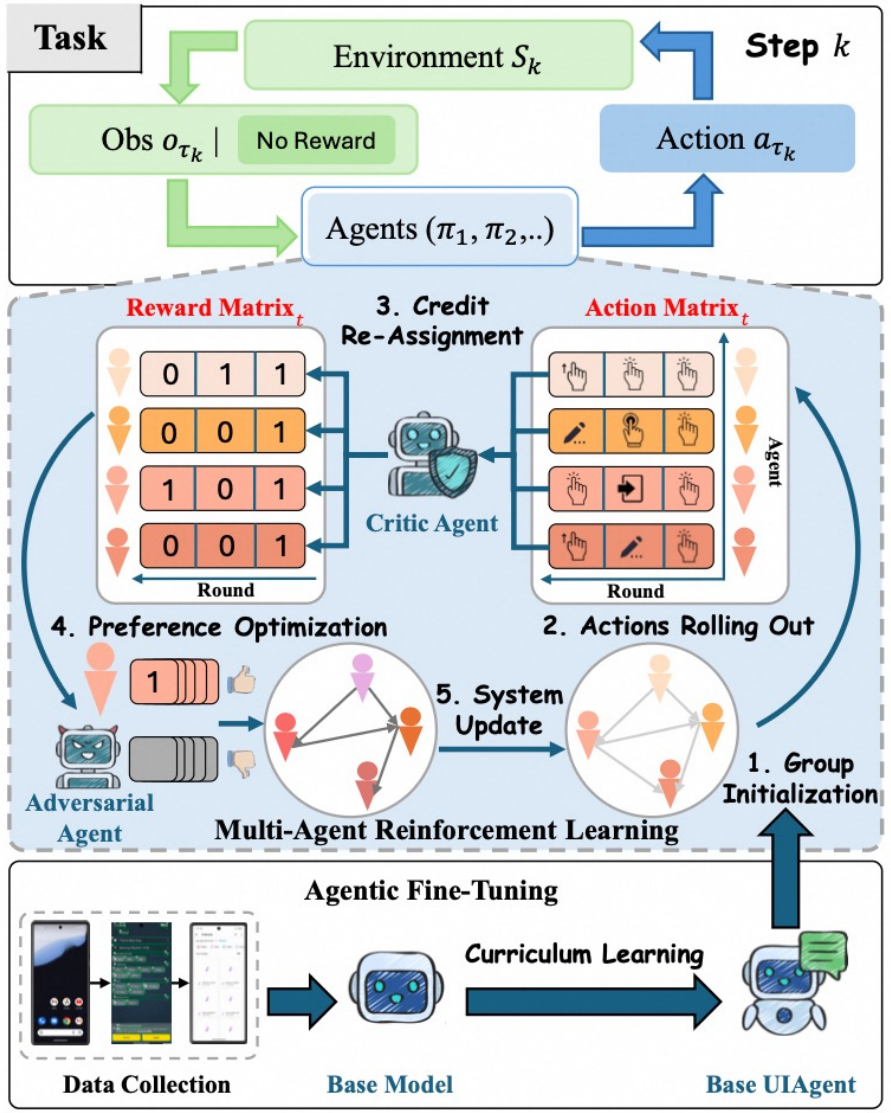
- **基于阶段一的智能体模型，构造多智能体策略系统，进而多智能体在线强化学习**

过程奖励分解策略

1. 对于每步决策，汇总每个智能体在多轮交流中的决策，得到**行为矩阵**
2. 评判智能体给出细粒度过程奖励，汇总为**奖励矩阵**

边更新策略

为防止训练过程中，多智能体过度依赖特定的多智体拓扑交流结构，在每轮训练后，随机**更新边集**至另一有向无环图，重组多智能体系统



He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

主要实验结果：移动终端UI环境训练与泛化结果

- **显著提升智能体在训练环境上的效果，同时有效泛化至其他移动终端UI环境**

(表示分布外环境效果提升值)

Method	Base model	#Params	#Agents	Input	$SR_{AndroidWorld}$	$SR_{MMiniWoB++}$	$\Delta_{Generalization}$
Agents based on Closed-Source LLMs							
M3A	GPT-4	N/A	1	Text	**30.6**	59.7	-
M3A	GPT-4	N/A	1	Text & Image	25.4	**67.7**	-
SeeAct	GPT-4	N/A	1	Text & Image	15.5	66.1	-
M3A	Gemini 1.5 Pro	N/A	1	Text	19.4	57.4	-
M3A	Gemini 1.5 Pro	N/A	1	Text & Image	22.8	40.3	-
Agents based on Open-Source LLMs							
Base Model	Qwen2	7B	1	Text	6.2	12.9	-
Base Model	Qwen2 VL	2B	1	Text & Image	0.0	10.0	-
InfiGUIAgent	Qwen2 VL	2B	1	Text & Image	9.1	15.6	5.6
Base Model	Qwen2.5 VL	3B	1	Text & Image	10.1	18.7	-
DigiRL	Qwen2.5 VL	3B	1	Text & Image	22.3	35.2	16.5
DigiRL	Qwen2.5 VL	3B	4	Text & Image	20.1	38.7	20.0
Our Methods							
Base UIAgent	Qwen2	7B	1	Text	18.9	48.4	35.5
Base UIAgent	Qwen2	7B	4	Text	21.4	53.2	40.3
$CollabUIAgents_{mobile}$	Qwen2	7B	4	Text	**29.3**	**61.2**	**48.3**

移动终端UI环境中，**仅在AndroidWorld环境上训练**，在训练与分布外环境上，显著强于基线强化学习方法，均接近或超过闭源模型单智能体效果

He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

主要实验结果：进一步扩展至网页UI环境

- **多智能体系统可直接应用于其他环境，或继续训练后扩展至其他环境**

- 移动终端UI环境多智能体系统**可直接应用于网页环境**，相较于开源模型智能体系统取得更好性能

System	#Params	#Agents	Input	Cross-Task	Cross-Website	Cross-Domain	Avg.
Agents based on Closed-Source LLMs							
GPT-3.5-Turbo	N/A	1	Text	17.4	16.2	18.6	17.4
GPT-4	N/A	1	Text	**36.2**	**30.1**	**26.4**	**30.9**
Agents based on Open-Source LLMs							
Qwen-VL*	9.6B	1	Text & Image	12.6	10.1	8.0	10.2
SeeClick*	9.6B	1	Text & Image	23.7	18.8	20.2	20.9
Qwen2	7B	1	Text	8.6	6.3	7.5	7.4
Our Methods							
Base UIAgent	7B	1	Text	13.4	10.6	11.8	11.9
Base UIAgent	7B	4	Text	15.7	11.2	12.9	13.2
CollabUIAgents$_{mobile}$	7B	4	Text	19.2	13.8	15.5	16.2
CollabUIAgents$_{m\rightarrow web}$ *	7B	4	Text	**34.5**	**32.7**	**25.1**	**30.7**

Mind2Web环境性能得分

System	#Params	#Agents	English		Chinese		Avg.
			Cross-Task	Cross-Domain	Cross-Task	Cross-Domain	
Agents based on Closed-Source LLMs							
GPT-3.5-Turbo	N/A	1	12.1	6.4	13.5	10.8	10.7
GPT-4	N/A	1	**38.6**	**39.7**	**36.7**	**36.3**	**37.8**
Claude2	N/A	1	13.2	8.1	13.0	7.9	10.5
Agents based on Open-Source LLMs							
LLaMA2	7B	1	3.3	2.5	-	-	2.9
LLaMA2	70B	1	8.3	8.9	-	-	10.6
Qwen2	7B	1	8.6	9.4	8.1	7.8	8.5
Our Methods							
Base UIAgent	7B	1	12.0	13.3	12.7	13.4	12.8
Base UIAgent	7B	4	13.7	14.5	15.0	13.9	14.0
CollabUIAgents$_{mobile}$	7B	4	18.6	17.7	19.1	15.6	17.7
CollabUIAgents$_{m\rightarrow web}$	7B	4	**34.3**	**36.9**	**35.3**	**32.5**	**34.7**

AutoWebBench性能得分

主要实验结果：进一步扩展至网页UI环境

- **多智能体系统可直接应用于其他环境，或继续训练后扩展至其他环境**

- 移动终端UI环境多智能体系统**可直接应用于网页环境**，相较于开源模型智能体系统取得更好性能

- 移动终端UI环境多智能体系统可在网页环境上**继续进行多智能体强化学习**，在训练环境（Mind2Web）和分布外环境（AutoWebBench）上，性能接近或超过闭源单智能体

System	#Params	#Agents	Input	Cross-Task	Cross-Website	Cross-Domain	Avg.
Agents based on Closed-Source LLMs							
GPT-3.5-Turbo	N/A	1	Text	17.4	16.2	18.6	17.4
GPT-4	N/A	1	Text	**36.2**	**30.1**	**26.4**	**30.9**
Agents based on Open-Source LLMs							
Qwen-VL*	9.6B	1	Text & Image	12.6	10.1	8.0	10.2
SeeClick*	9.6B	1	Text & Image	23.7	18.8	20.2	20.9
Qwen2	7B	1	Text	8.6	6.3	7.5	7.4
Our Methods							
Base UIAgent	7B	1	Text	13.4	10.6	11.8	11.9
Base UIAgent	7B	4	Text	15.7	11.2	12.9	13.2
CollabUIAgents$_{mobile}$	7B	4	Text	19.2	13.8	15.5	16.2
CollabUIAgents$_{m\rightarrow web}$ *	7B	4	Text	**34.5**	**32.7**	**25.1**	**30.7**

Mind2Web环境性能得分

System	#Params	#Agents	English		Chinese		Avg.
			Cross-Task	Cross-Domain	Cross-Task	Cross-Domain	
Agents based on Closed-Source LLMs							
GPT-3.5-Turbo	N/A	1	12.1	6.4	13.5	10.8	10.7
GPT-4	N/A	1	**38.6**	**39.7**	**36.7**	**36.3**	**37.8**
Claude2	N/A	1	13.2	8.1	13.0	7.9	10.5
Agents based on Open-Source LLMs							
LLaMA2	7B	1	3.3	2.5	-	-	2.9
LLaMA2	70B	1	8.3	8.9	-	-	10.6
Qwen2	7B	1	8.6	9.4	8.1	7.8	8.5
Our Methods							
Base UIAgent	7B	1	12.0	13.3	12.7	13.4	12.8
Base UIAgent	7B	4	13.7	14.5	15.0	13.9	14.0
CollabUIAgents$_{mobile}$	7B	4	18.6	17.7	19.1	15.6	17.7
CollabUIAgents$_{m\rightarrow web}$	7B	4	**34.3**	**36.9**	**35.3**	**32.5**	**34.7**

AutoWebBench性能得分

消融实验结果

- 通过多智能体强化学习进行环境扩展，对原本训练环境上的性能影响不大

Method	$SR_{AndroidWorld}$	$SR_{MMiniWoB++}$
Single-Agent Systems		
Qwen2	6.2	12.9
Base UIAgent	18.9	48.4
Multi-Agent Systems ($n = 4$)		
Qwen2	8.6	16.1
Base UIAgent	21.4	53.2
$CollabUIAgents_{mobile}$	**29.3**	**61.2**
w/ PO → RFT	23.2	54.8
w/o CR	25.0	56.4
w/o Edge Update	27.6	58.1
$CollabUIAgents_{m\rightarrow web}$	26.7	58.1

移动终端UI环境性能得分

He et al. Enhancing Language Multi-Agent Learning with Multi-Agent Credit Re-Assignment for Interactive Environment Generalization. arXiv:2502.14496.

消融实验结果

- **通过多智能体强化学习进行环境扩展，对原本训练环境上的性能影响不大**

Method	$SR_{AndroidWorld}$	$SR_{MMiniWoB++}$
Single-Agent Systems		
Qwen2	6.2	12.9
Base UIAgent	18.9	48.4
Multi-Agent Systems ($n = 4$)		
Qwen2	8.6	16.1
Base UIAgent	21.4	53.2
$CollabUIAgents_{mobile}$	**29.3**	**61.2**
w/ PO → RFT	23.2	54.8
w/o CR	25.0	56.4
w/o Edge Update	27.6	58.1
$CollabUIAgents_{m \rightarrow web}$	26.7	58.1

移动终端UI环境性能得分

其他一些insights：

- 同一基础模型上，合理设计的多智能体系统一般强于单智能体；但现有开源的通用指令模型/单智能体模型，对多智能体系统的适应不佳，**瓶颈可能在于模型多智体能力**

消融实验结果

- **通过多智能体强化学习进行环境扩展，对原本训练环境上的性能影响不大**

Method	$SR_{AndroidWorld}$	$SR_{MMiniWoB++}$
Single-Agent Systems		
Qwen2	6.2	12.9
Base UIAgent	18.9	48.4
Multi-Agent Systems ($n = 4$)		
Qwen2	8.6	16.1
Base UIAgent	21.4	53.2
$CollabUIAgents_{mobile}$	**29.3**	**61.2**
w/ PO → RFT	23.2	54.8
w/o CR	25.0	56.4
w/o Edge Update	27.6	58.1
$CollabUIAgents_{m \to web}$	26.7	58.1

移动终端UI环境性能得分

其他一些insights：

- 现有开源的通用指令模型/单智能体模型，对多智能体系统的适应不佳，**瓶颈可能在于模型多智体能力**
- 复杂环境中，单/多智能体训练时，**强化学习的效率在经验上强于有监督的指令微调（RFT）**，可能是获取完全正确的训练轨迹比较困难

未来可能的探索方向

1. 基于多模态模型（主要是VLM）的多智能体系统架构与训练方法

- 多模态多智能体系统涉及较复杂的通信协议设计；
- 多模态模型的多智能体强化学习空间较大，信号更加稀疏平凡，需要进一步探索。

2. 基于现有前沿多智能体框架（例如Mobile-Agent v2/v3）的适配

- 若直接使用当前训练所得模型，效果不稳定；
- 当前方法可以接入其他多智能体系统架构，但需要重新训练。

3. 基于长推理模型（o1/3, R1, ...）的智能体的适配与训练

- 智能体响应长推理时，难以满足延迟、环境交互效率等要求；
- 可能需要选择性执行长推理（例如，将<think>作为工具），并纳入在线强化学习框架。

……