北京郵電大學
Beijing University of Posts and Telecommunications

AgentRefine: Enhancing Agent Generalization through Refinement Tuning

Dayuan Fu[1*],Keqing He[2*],Yejie Wang[1*],Wentao Hong[1], Zhuoma Gongque[1], Weihao Zeng[1], Wei Wang[2], Jingang Wang[2], Xunliang Cai[2], Weiran Xu[1 †]

[1]Beijing University of Posts and Telecommunications

[2] Meituan, Beijing, China

[*]Equal contribution. Emails: fdy@bupt.edu.cn

[†] Corresponding authors.

Content

1 Motivations
2 Contribution
3 Methodology
4 Experiments
5 Analysis
6 Conclusion

Motivations

- Current agent training corpora perform well on held-in evaluation sets but show poor generalization on held-out evaluation sets.
- Agent tuning efforts suffer from severe formatting errors, often repeating the same mistakes for extended periods. They cannot learn from experience and are limited to memorizing existing observation-action relationships.

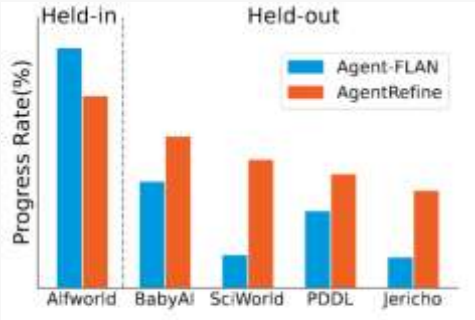

Figure 1: Overall progress score among 5 tasks. Agent-FLAN has been trained on Held-in task.

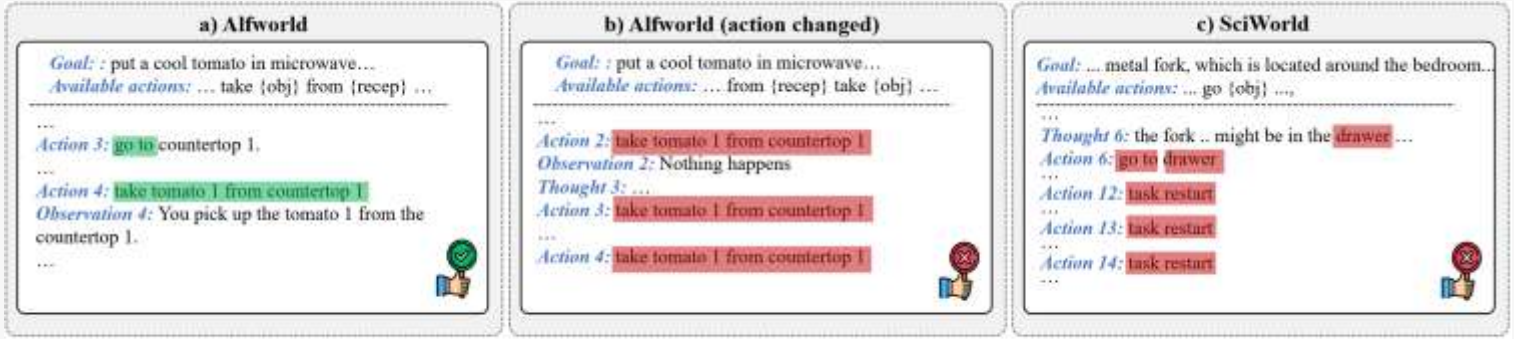

Figure 2: Example of parameter memorization in Agent-FLAN.

Contributions

We link agent generalization to self-refinement using feedback from the environment. A capable agent should recognize mistakes, refine actions, and learn from errors to avoid repetitive pitfalls and discover correct action sequences.

- While existing agent-tuning work improve held-in agent performance, they hardly generalizes the ability to new agent tasks. In contrast, our AgentRefine does not depend on memorizing training trajectories but learns to self-refine its mistakes and explore more actions and reasonable paths
- Our experiments demonstrate that agent-tuning on normal trajectories performs poorly to the small perturbation of agent environments, like the action description. Refinement tuning exhibits greater robustness to environmental changes.
- Further analysis indicates the diversity of agent environments and thoughts contributes to refinement tuning.

Methodology

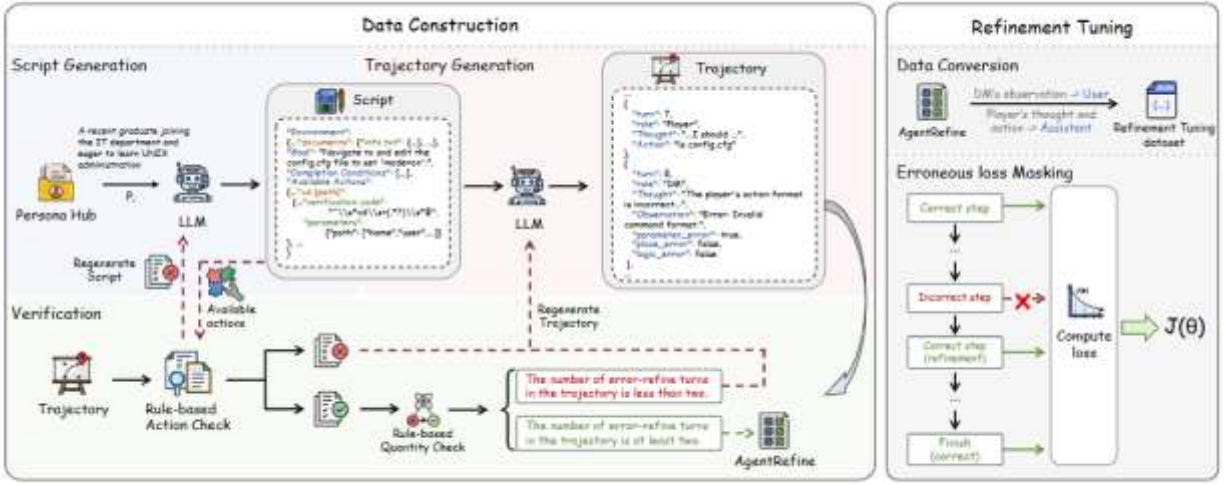

Figure 4: The pipeline of AgentRefine data generation and refinement tuning.

- Data Construction:

step1. Script Generation
step2. Trajectory Generation
step3. Verification

- Training

Mask incorrect step and use Refinement tuning.

Experiment

- Main Results

Method	Alfworld		BabyAI		SciWorld		PDDL		Jericho	
	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress
GPT Series										
GPT-4o	66.4	79.9	48.2	64.1	40	76.9	61.7	69.8	10.0	34.0
GPT-4o-mini	37.3	65.0	36.6	51.9	23.3	49.8	25.0	49.1	10.0	28.5
LLaMA-3-8B Series										
LLaMA-3-8B-Instruct	22.4	46.1	45.5	56.5	7.8	41.1	10.0	38.4	0.0	24.3
AgentGen	29.1	47.6	20.5	35.0	-	-	11.7	23.0	-	-
AgentGym	61.9	76.9	47.3	61.4	18.9	47.5	1.7	16.6	0.0	12.9
Agent-FLAN	67.2	79.7	25.0	35.3	1.1	10.9	8.3	25.5	0.0	10.1
AgentRefine	44.8	63.8	37.5	50.4	14.4	42.6	16.6	37.8	10.0	32.3
Mistral Series										
Mistral-7B-Instruct-v0.3	12.4	35.9	36.6	45.8	6.7	24.7	13.3	27.8	0.0	17.3
AgentGym	76.9	86.7	40.2	56.3	15.6	48.3	1.7	7.3	0.0	13.0
Agent-FLAN	77.6	87.6	15.2	21.0	0	6.7	0	3.2	0.0	0.7
AgentRefine	51.4	68.8	25.9	42.4	4.4	22.4	11.7	32.8	5.0	28.8
LLaMA-3-70B Series										
LLaMA-3-70B-Instruct	67.2	75.2	48.2	61.8	42.2	75.4	55.0	79.8	25.0	46.4
Agent-FLAN	80.5	86.8	32.1	41.2	5.5	16.4	25.0	53.7	0.0	13.6
AgentRefine	67.2	72.1	44.6	59.7	17.7	46.4	38.3	58.6	15.0	37.2

Table 1: Main Results. The underlined text indicates that the training data is sampled in the same environment as the task and is considered as held-in evaluation. We use the original result in AgentGen and reproduce AgentGym and Agent-FLAN's results.

Analysis

- Ablation study

Method	Alfworld		BabyAI		SciWorld		PDDL		Jericho	
	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress
AgentRefine	48.5	61.5	37.1	51.7	7.7	33.1	21.7	37.4	5.0	26.2
- w/o refinement loss	40.3	58.8	34.8	45.6	4.4	22.7	20.0	37.4	0.0	16.1
- w/o refinement data	49.3	65.2	30.4	43.1	5.5	21.3	11.7	32.5	0.0	13.8
- w erroneous loss	29.9	43.9	23.2	31.6	3.3	19.0	8.3	28.3	5.0	18.4

Table 2: Ablation study of Refinement Tuning. This experiment is in the data size of 8000.

- Robustness Analysis

Model	Alfworld		Perturbation 1		Perturbation 2		Perturbation 3		Perturbation 4		Perturbation 5		Average		Std	
	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress
LLaMA3-8B-Instruct	22.4	46.1	23.1	45.6	24.6	45.0	17.9	45.1	17.9	45.1	22.4	46.1	21.4	45.5	2.68	0.47
AgentGym	61.9	76.9	29.1	59.2	49.2	65.3	32.8	53.9	38.8	48.2	5.9	28.7	36.3	55.4	19.97	16.66
Agent-FLAN	67.2	79.7	21.6	58.8	51.4	71.3	27.6	53.5	52.2	67.9	1.5	19.7	36.9	58.5	21.98	22.53
AgentRefine	44.8	63.8	50.0	66.5	51.5	66.7	54.5	70.0	45.5	60.6	44.8	63.8	48.5	65.2	3.73	3.56

Table 3: Performance for different models across various perturbations.

Analysis

- Diversity analysis

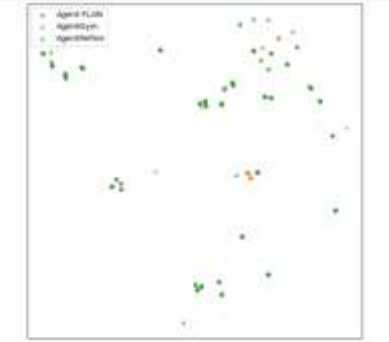

Figure 6: The t-SNE figure among Agent-FLAN, AgentGym, and AgentRefine's Thought.

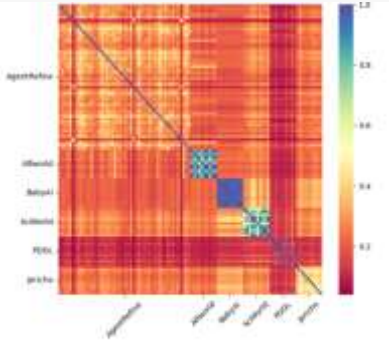

Figure 7: The similarity heatmap between different environments in 6 sources.

Model	Alfworld		BabyAI		SciWorld		PDDL		Jericho	
	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress
AgentGym-greedy	61.9	76.9	47.3	61.4	18.9	47.5	1.7	16.6	0.0	12.9
AgentGym-BoN	99.3	99.3	73.2	87.2	58.9	85.6	16.6	42.1	5.0	22.2
Δ	37.4	22.4	25.9	25.8	40.0	38.1	14.9	25.5	5.0	9.3
Agent-FLAN-greedy	67.2	79.7	25.0	35.3	1.1	10.9	8.3	25.5	0.0	10.1
Agent-FLAN-BoN	85.5	98.1	43.8	56.7	10.0	33.5	11.7	39.8	5.0	22.2
Δ	28.3	18.4	18.8	21.4	8.9	22.6	3.4	14.3	5.0	12.1
AgentRefine-greedy	44.8	63.8	37.5	50.4	14.4	42.6	16.6	37.8	10.0	32.3
AgentRefine-BoN	93.3	96.6	67.0	81.5	40.0	71.0	30.0	57.3	25	52.5
Δ	48.5	32.8	29.5	31.1	25.6	28.4	13.4	19.5	15.0	20.2

Table 4: Best-of-N results among three methods.

Analysis

- Synthesis from open source model

Model	Alfworld		BabyAI		SciWorld		PDDL		Jericho	
	Success	Progress	Success	Progress	Success	Progress	Success	Progress	Success	Progress
Agent-FLAN	67.2	79.7	25.0	35.3	1.1	10.9	8.3	25.5	0.0	10.1
AgentRefine-DeepSeek	32.0	44.2	36.6	48.1	2.2	21.6	16.6	36.7	5.0	29.0
AgentRefine-GPT-4o	36.6	55.9	33.9	44.1	11.1	31.4	18.3	37.9	10.0	28.8

Table 5: Performance on Different Synthesis Models, we synthesize 4000 data via deepseek-v2.5. The underlined text indicates that the training data is sampled in the same environment as the task and is considered as held-in evaluation

- Generalization in reasoning task

Method	EM	F1
LLaMA-3-8B-Instruct	29.3	36.6
AgentGym	28.0	37.4
Agent-FLAN	24.6	32.4
AgentRefine	37.0	44.6

Figure 8: Model Performance on reasoning task, Hotpot QA.

Analysis

- Case study

Figure 9: Comparison case study on Jericho and SciWorld between Agent-FLAN and AgentRefine.

Analysis

- GPT-4 Judgement's reliability

Human \ GPT-4	Right	Wrong
Right	47	9
Wrong	3	41

Figure 10: The comparison of GPT-4's judgement and human's judgement. The right column/line means human/GPT-4 considers this turn doesn't need to be refined. The wrong column/line means human/GPT-4 considers this turn needs to be refined.

- Generalization between general data and agent data

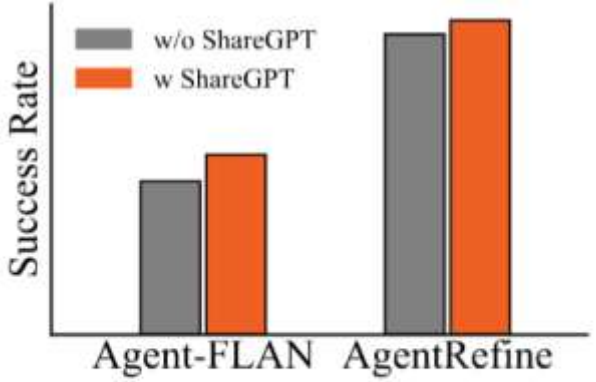

Figure 11: The success rate by incorporating ShareGPT

Conclusion

In this work, we study the generalized agent abilities for open-source LLMs via agent tuning. Current work performs well on held-in evaluation sets but fails to generalize to held-out sets because of overfitting to several manual agent environments. We present the AgentRefine approach to enable the model to correct its mistakes based on the environment feedback. Experiments demonstrate that AgentRefine significantly outperforms state-of-the-art agent-tuning work in terms of generalization ability on diverse agent benchmarks. Our analysis shows that self-refinement enables the robustness of agent capability and the diversity of agent environments and thoughts further enhances the performance. We hope to provide new insight for future agent research.

THANKS!