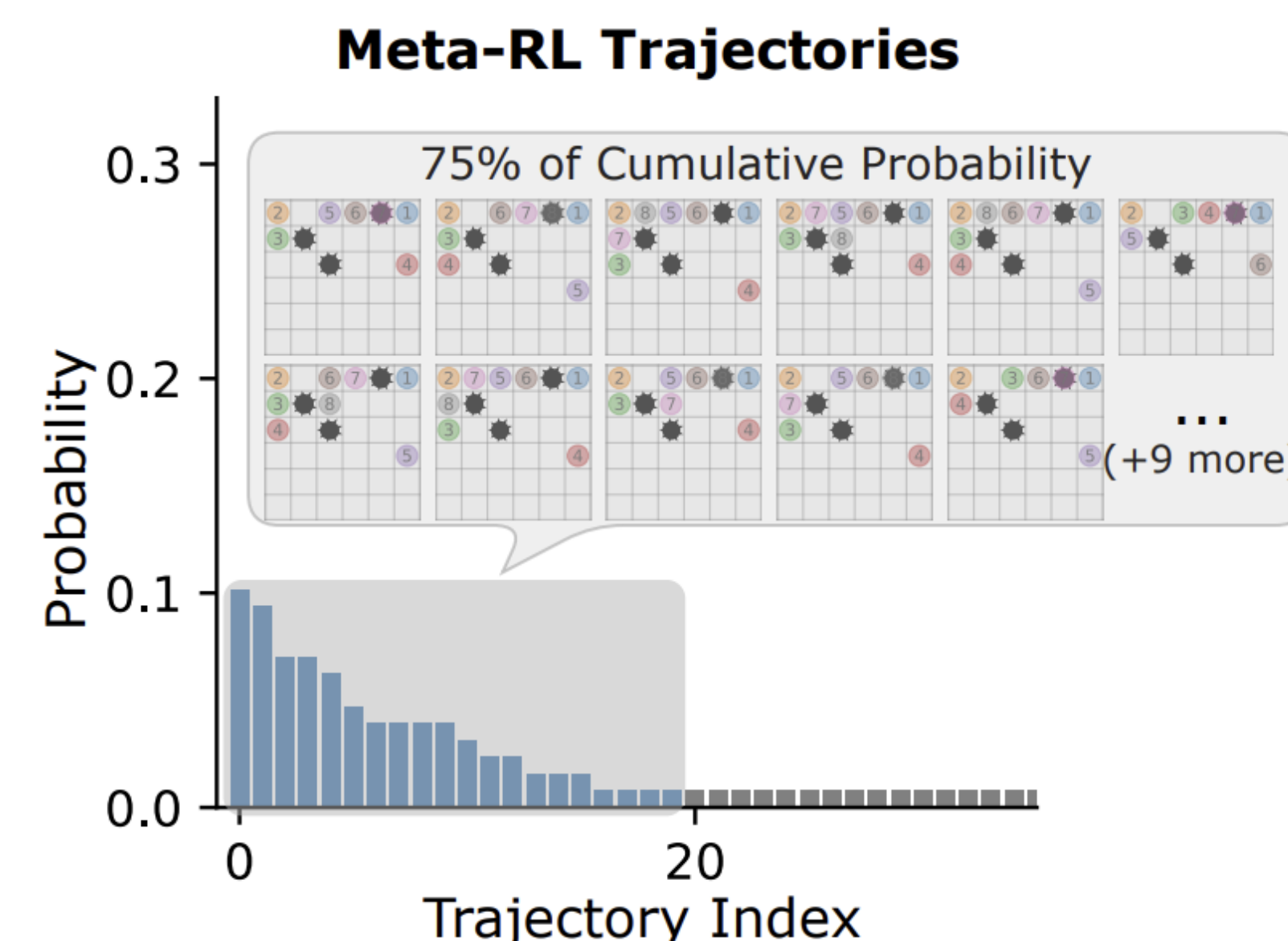
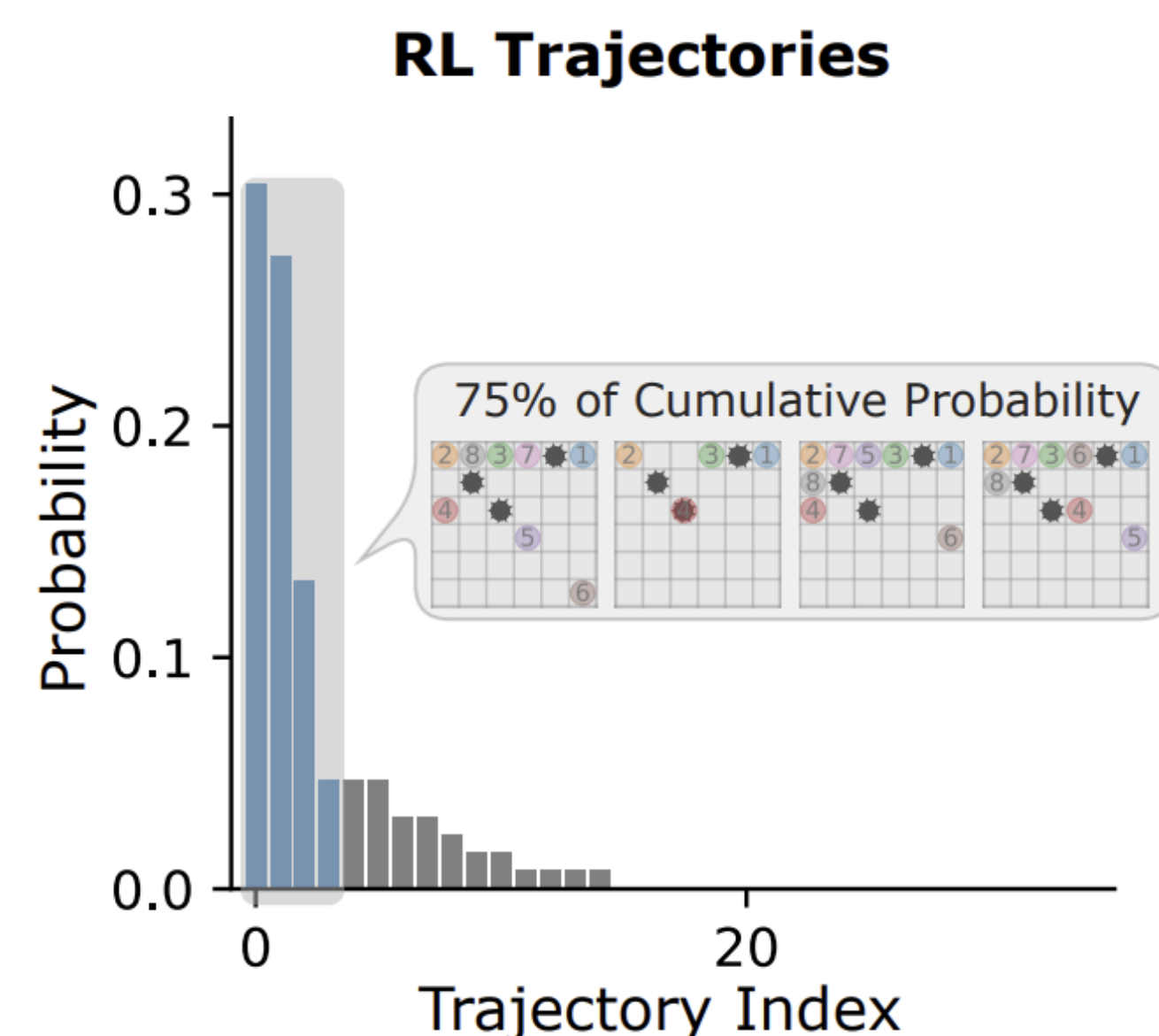
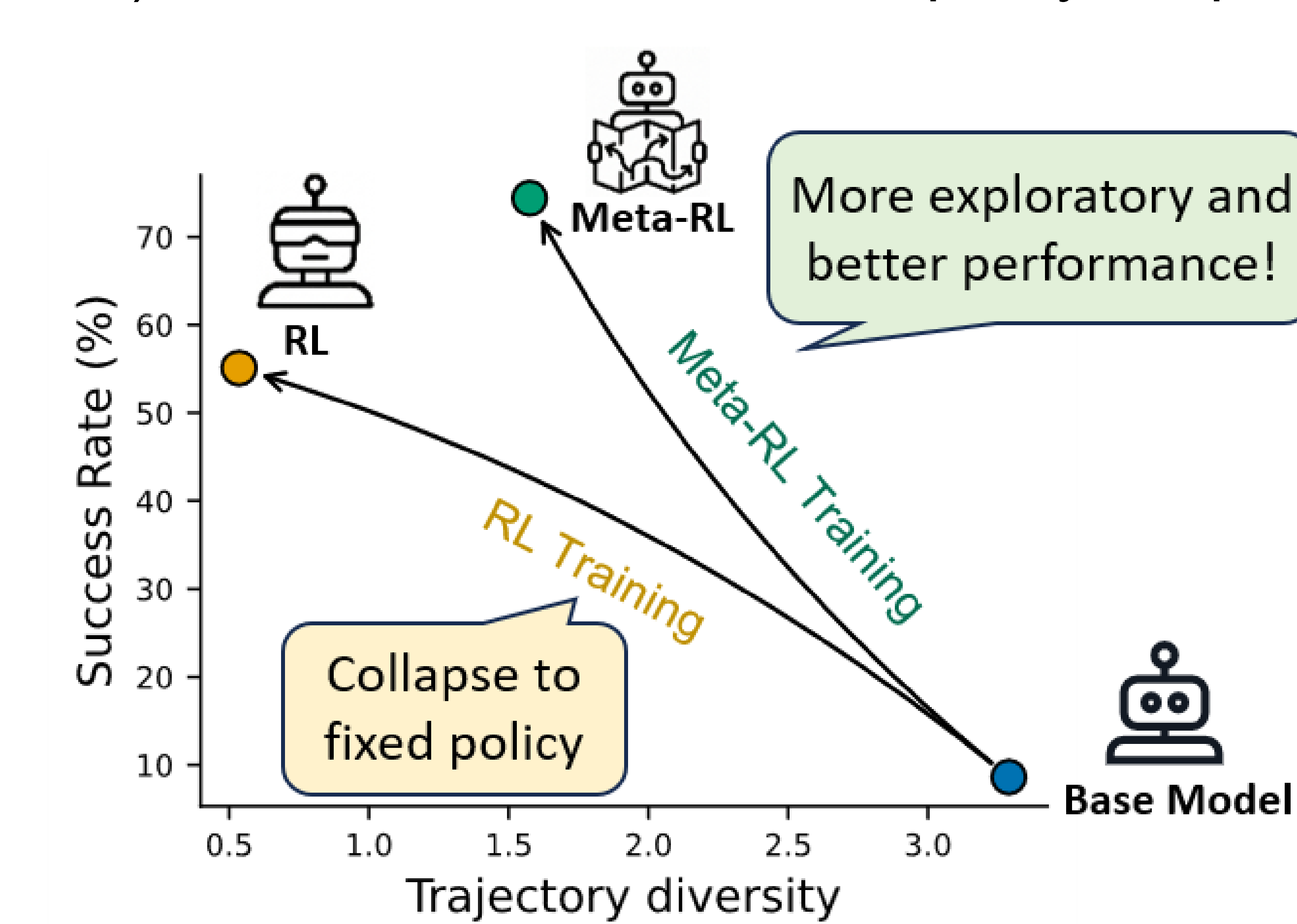


Meta-RL Induces Exploration in Language Agents

Yulun Jiang*, Liangze Jiang*, Damien Teney, Michael Moor†, Maria Brbić†

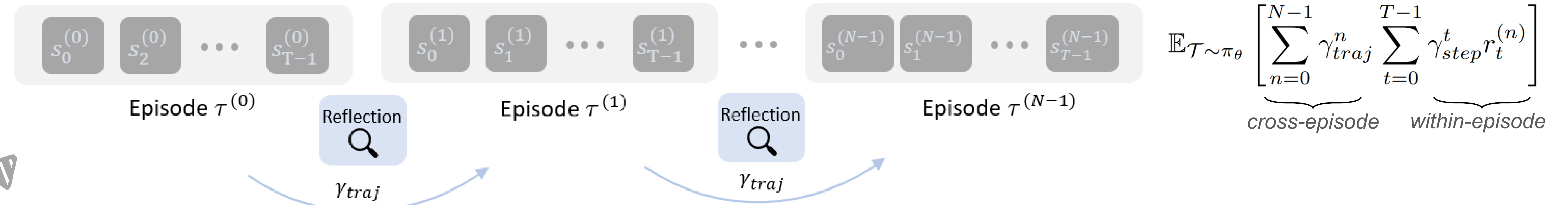
Takeaways

- LLM agents often struggle in **multi-turn** tasks that require **active exploration**.
- **LaMer** induces exploration-exploitation via
 - 1) cross-episode **Meta-RL** training;
 - 2) reflection-based in-context policy adaptation.

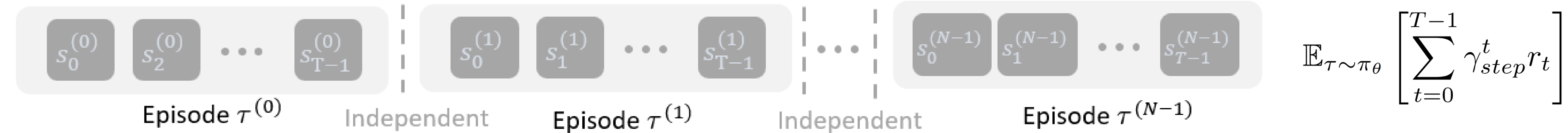


LLM Agent with Meta-RL (LaMer)

- **LaMer** trains agents across multiple episodes, maximizing a discounted cross-episode return.



- **Standard RL training** maximizes single-episode return.

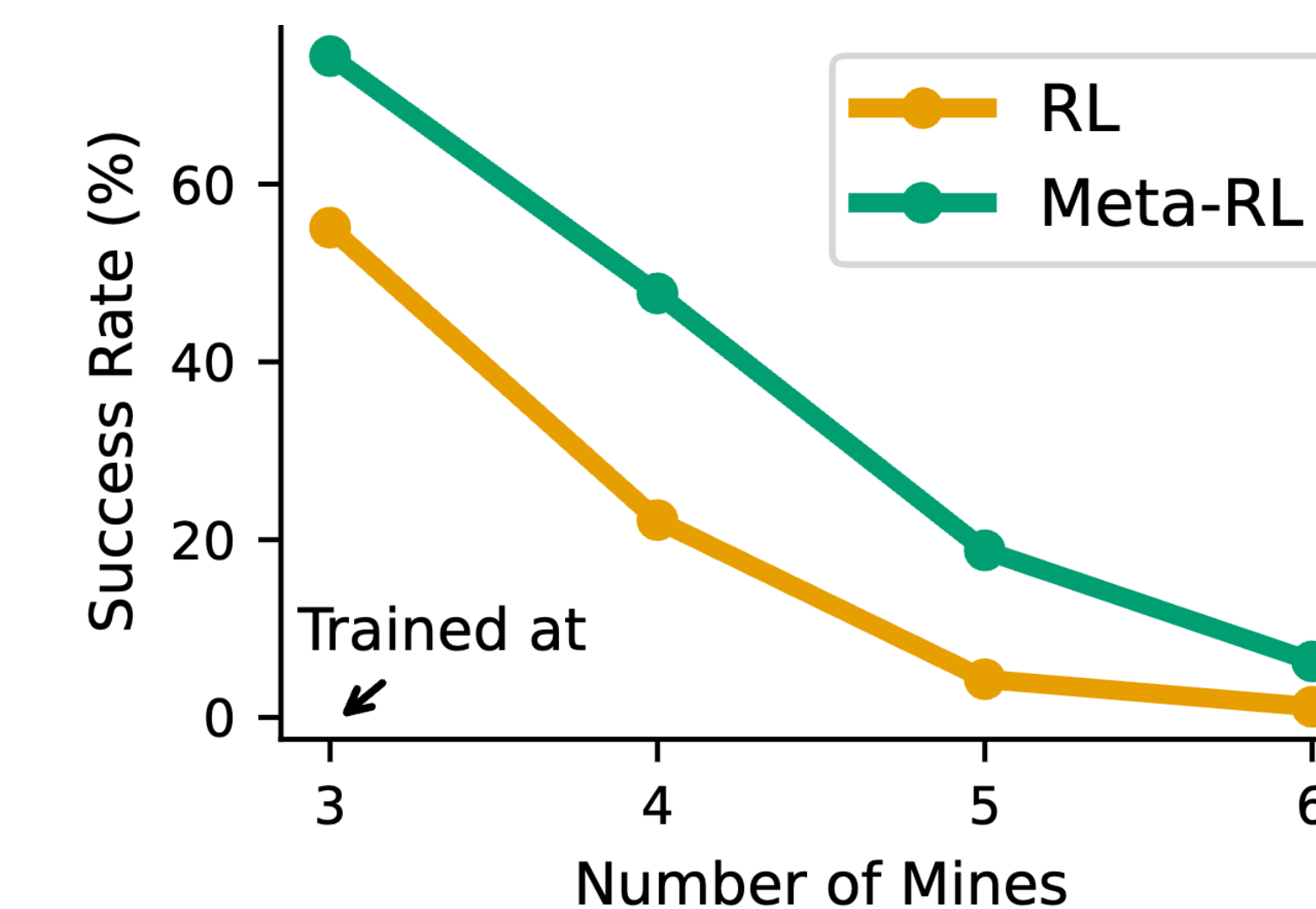


Key Results

- ✓ Unlock strong performance and test-time scaling

Method	Sokoban			MineSweeper			Webshop		
	p@1	p@2	p@3	p@1	p@2	p@3	p@1	p@2	p@3
<i>Prompting</i>									
Zero-shot	6.8	9.8	12.9	4.5	6.6	8.6	1.4	2.1	2.3
ReAct	7.2	9.6	12.5	6.3	7.0	10.9	3.1	4.5	4.5
Reflexion	6.4	9.8	12.1	5.5	7.2	9.8	2.7	3.3	3.5
<i>Training with RL</i>									
PPO	12.5	15.4	16.8	29.7	34.2	35.5	53.1	54.5	54.9
RLOO	13.5	16.6	18.8	48.8	51.2	51.6	67.6	68.4	69.1
GRPO	22.9	26.4	27.0	36.3	40.0	40.4	72.9	73.0	73.0
GiGPO	41.6	43.6	44.1	52.0	54.9	55.1	73.4	74.6	75.2
<i>Training with Meta-RL (ours)</i>									
LAMER	42.4	52.0	55.9	44.1	66.4	74.4	67.8	84.4	89.1

- ✓ Generalize to harder or out-of-distribution tasks



(MineSweeper)

Method	i.d				o.o.d	
	Pick	Look	Clean	Heat	Cool	Pick2
Prompting	91.9	52.9	48.4	44.8	42.8	21.2
RL	95.5	83.0	67.9	86.6	58.1	36.0
Meta-RL	97.7	100.0	90.2	89.5	81.0	50.2

(ALFWorld)

★ ...more analyses & ablations in the paper!