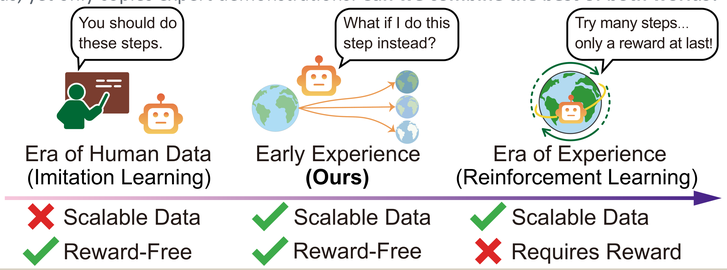
Agent Learning via *Early Experience*

Kai Zhang, Xiangchao Chen, Bo Liu, Tianci Xue, Zeyi Liao, Zhihan Liu, Xiyao Wang, Yuting Ning, Zhaorun Chen, Xiaohan Fu, Jian Xie, Yuxuan Sun, Boyu Gou, Qi Qi, Zihang Meng, Jianwei Yang, Ning Zhang,
Xian Li, Ashish Shah, Dat Huynh, Hengduo Li, Zi Yang, Sara Cao, Lawrence Jang, Shuyan Zhou, Jiacheng Zhu, Huan Sun, Jason Weston, Yu Su†, Yifan Wu†

A reward-free bridge from imitation learning to RL

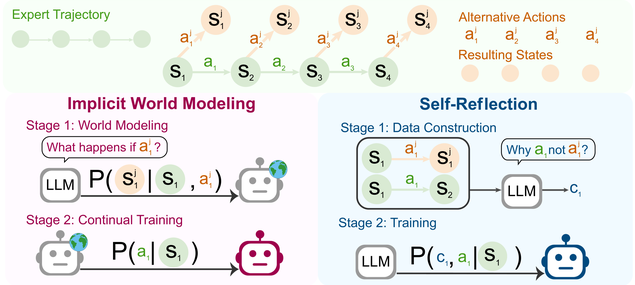
1 Motivation

Language agents should improve from their own experience. But RL needs rewards that many real environments — websites, multi-turn tools — rarely provide. Imitation learning (SFT) avoids rewards, yet only copies expert demonstrations. **Can we combine the best of both worlds?**

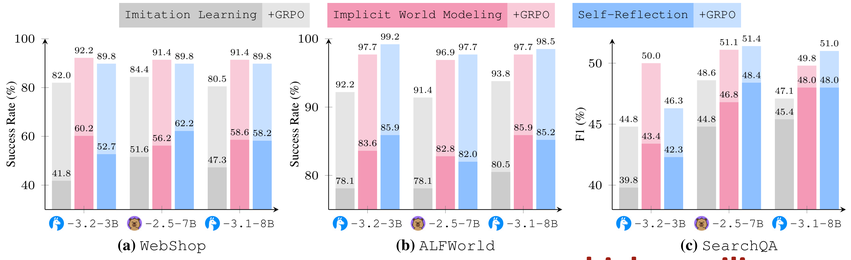


2 Early Experience

The agent proposes its own actions and collects the resulting future states as supervision. Two strategies turn these states into learning signals:

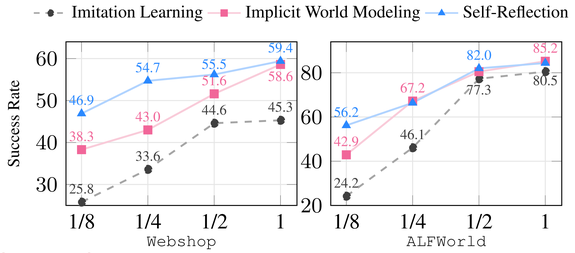


3 Results — Better Foundations for Upcoming RL, Data Efficiency, and Domain Generalization



(a) WebShop (b) ALFWorld (c) SearchQA

RL from early-experience checkpoints reaches **higher ceilings**



1/8 – 1/2 data w/ early experience matches full-data IL

	ALFWorld			BFCLv3			SearchQA (F1)		
	-3.2-3B	-2.5-7B	-3.1-8B	-3.2-3B	-2.5-7B	-3.1-8B	-3.2-3B	-2.5-7B	-3.1-8B
Prompt	5.5	4.7	18.8	1.3	7.1	6.2	24.6	33.1	37.0
Imitation Learning	74.2	64.1	63.3	5.3	7.6	6.7	40.5	47.0	47.4
Ours-IWM	**77.3** (+3.1)	70.3 (+6.2)	**78.1** (+14.8)	8.9 (+3.6)	**12.9** (+5.3)	7.6 (+0.9)	**45.4** (+4.9)	49.5 (+2.5)	49.6 (+2.2)
Ours-SR	**77.3** (+3.1)	**71.1** (+7.0)	72.7 (+9.4)	**13.8** (+8.5)	8.3 (+0.7)	**8.0** (+1.3)	44.0 (+3.5)	**51.2** (+4.2)	**50.7** (+3.3)

Out-of-domain gains persist across three unseen splits— even **larger than in-domain**

3 Results — Immediate Gains to IL

Benchmark	Model	Prompt	IL	Ours-IWM	Ours-SR
Embodied and Scientific Simulation, and Travel Planning					
ALFWorld	-3.2-3B	8.6	78.1	83.6 (+5.5)	**85.9** (+7.8)
	-2.5-7B	14.8	78.1	**82.8** (+4.7)	82.0 (+3.9)
	-3.1-8B	25.0	80.5	**85.9** (+5.4)	85.2 (+4.7)
ScienceWorld	-3.2-3B	2.3	51.6	55.5 (+3.9)	**56.2** (+4.6)
	-2.5-7B	3.9	53.9	**59.4** (+5.5)	57.8 (+3.9)
	-3.1-8B	3.1	54.7	57.0 (+2.3)	**68.0** (+13.3)
TravelPlanner	-3.2-3B	0.0	19.4	28.3 (+8.9)	**32.2** (+12.8)
	-2.5-7B	0.0	16.7	22.2 (+5.5)	**31.7** (+15.0)
	-3.1-8B	0.0	17.2	25.0 (+7.8)	**32.2** (+15.0)
Multi-Turn Tool Use					
BFCLv3	-3.2-3B	1.3	21.3	25.3 (+4.0)	**29.3** (+8.0)
	-2.5-7B	10.6	26.7	29.3 (+2.6)	**32.0** (+5.3)
	-3.1-8B	6.7	16.0	**20.0** (+4.0)	**20.0** (+4.0)
Tau-Bench	-3.2-3B	5.2	24.3	26.1 (+1.8)	**28.7** (+4.4)
	-2.5-7B	20.0	33.9	38.7 (+4.8)	**39.5** (+5.6)
	-3.1-8B	6.0	35.9	40.8 (+4.9)	**41.7** (+5.8)
SearchQA (F1)	-3.2-3B	13.3	38.0	**39.0** (+1.0)	38.6 (+0.6)
	-2.5-7B	19.3	39.9	40.8 (+0.9)	**42.0** (+2.1)
	-3.1-8B	21.0	41.0	**44.3** (+3.3)	41.8 (+0.8)
Web Navigation					
WebShop	-3.2-3B	0.0	41.8	**60.2** (+18.4)	52.7 (+10.9)
	-2.5-7B	0.8	51.6	56.2 (+4.6)	**62.2** (+10.6)
	-3.1-8B	0.0	47.3	**58.6** (+11.3)	58.2 (+10.9)
WebArena-Lite	-3.2-3B	1.2	6.1	**8.5** (+2.4)	7.3 (+1.2)
	-2.5-7B	1.8	4.2	**7.3** (+3.1)	6.1 (+1.9)
	-3.1-8B	0.6	4.9	**8.5** (+3.6)	**8.5** (+3.6)

Consistently improves across **3** models and **8** environments