Agent Learning via Early Experience

2025.11.18

가짜연구소 박지예

https://arxiv.org/abs/2510.08558
https://huggingface.co/papers/2509.07979

PseudoLab

Introduction

- 대규모의 사전학습과 언어적으로 유연한 인터페이스로, 실제 환경에서 작업 가능한 에이전트가 현실화되고 있음
  - 웹 네비게이션, 모바일 조작, 연구 보조 등의 업무 수행 가능해짐
  - LLM이 시각적 문맥을 이해하고 다양한 과제를 수행 가능
- 에이전트의 학습에서는 보통 RL(Reinforcement Learning)을 통해 학습
  - RL은 환경에서 얻는 보상을 통해, 에이전트가 더 나아진 행동을 학습할 수 있게 함
  - 하지만 **RL은 정확한 보상 신호가 있어야 학습 가능하기 때문에, 명확한 보상 구조가 없는 실제 환경에서는 적용이 힘듦**

    [예시] 웹에서 폼을 제출하는 행동에 대해, 제출 버튼을 눌러도 성공 여부를 자동으로 판단하기 어려움

    (실제로 폼의 내용이 잘 채워졌는지 시스템이 알 수 없음 → 보상 X)
  - 또한 **여러 도구를 순차적으로 써야하는 multi-tool 환경처럼 행동 길이가 길면, RL의 학습이 불안정해짐**

Introduction

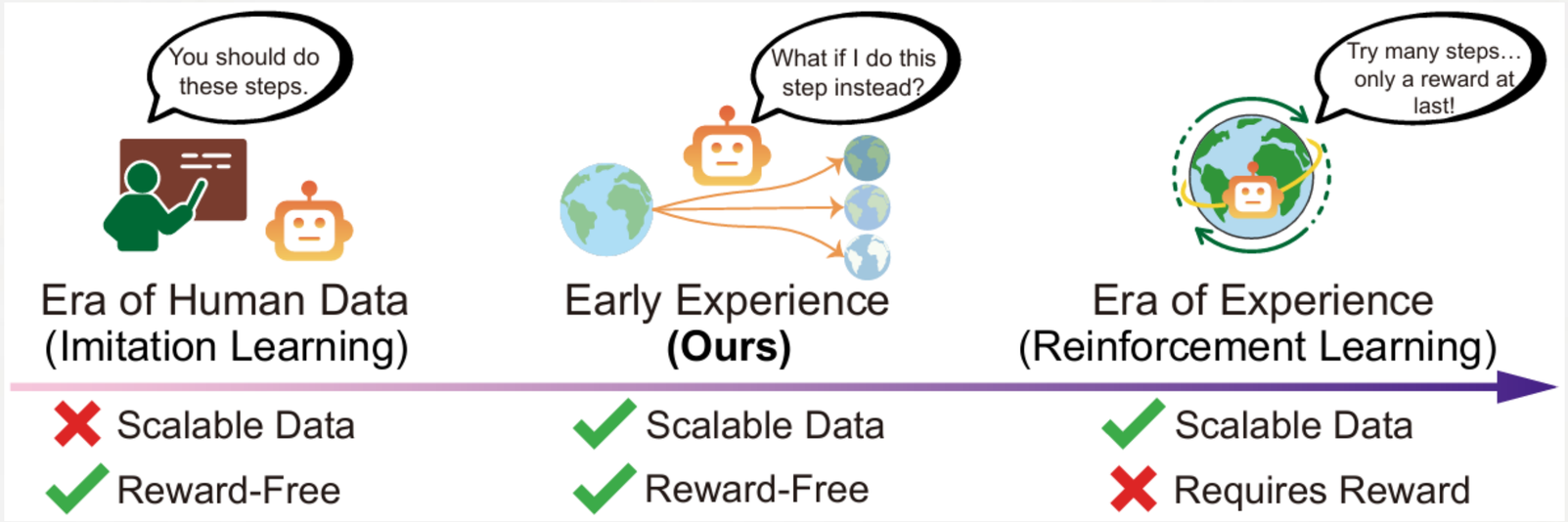
- RL의 여러 문제 때문에, 실제 환경에서는 SFT(Supervised Fine-Tuning)를 적용함
  - SFT는 보상 구조가 필요하지 않고, 사람이 만든 데모로 학습하기 때문에 안정적임
  - 또한 정적 시나리오 기반의 데이터로 데이터 수집이 용이하며, 구현 난이도가 낮음
- 에이전트 학습 시, **SFT는 실제의 환경적인 변수들을 고려하지 못한다는 한계**가 존재함
  - RL은 환경에서 얻는 보상을 통해, 에이전트가 더 나아진 행동을 학습할 수 있게 함
  - **에이전트가 자신의 행동에 대한 결과를 경험하지 못하고, 새로운 상황에 대해 취약함** → 개선 불가(모델이 수동적)

외부의 보상 없이도, 경험을 통해 발전 가능한 에이전트를 만드는 방법 필요

Introduction

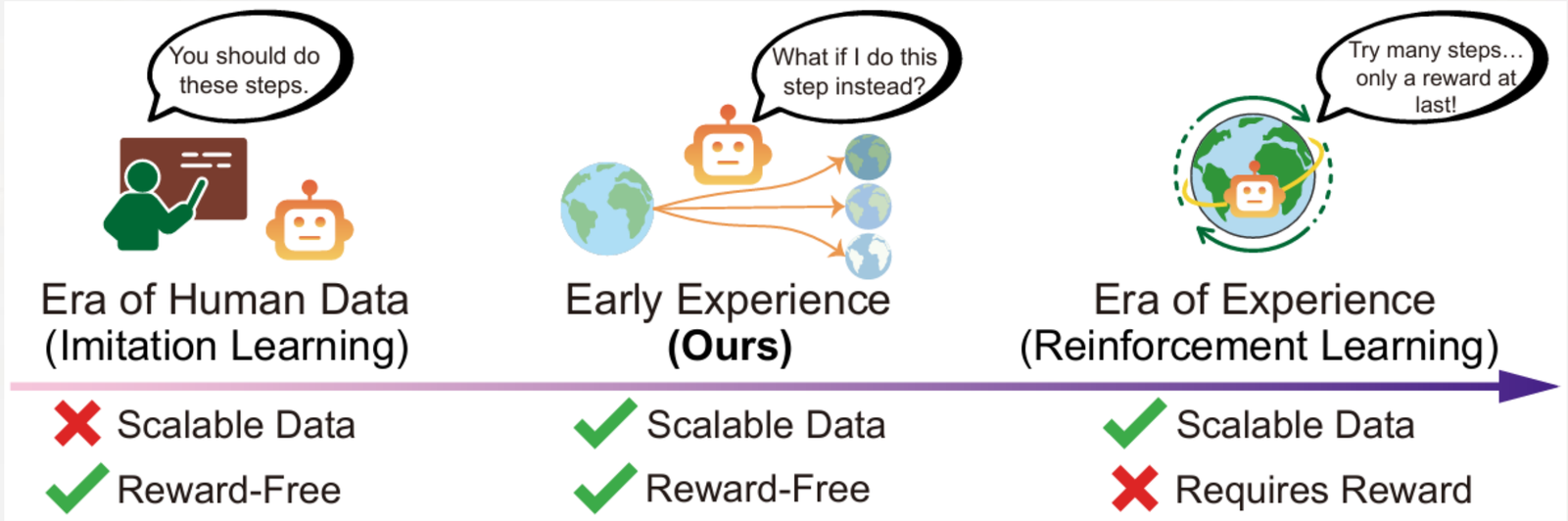
- **Early Experience**는 에이전트가 실제 수행하게 될 행동들의 미래 상태를 수집해서, 이 환경 변화를 학습 신호로 삼는 self supervised 기반의 학습 방식
  - 환경을 직접 변화시키면서 나오는 **미래 상태를 학습의 supervision으로 활용함**
  - 미래 상태가 행동의 결과이기 때문에, 다른 보상이 없어도 학습 신호를 통해 보상으로 활용 가능함
- Proposed Method
  - **Implicit World Modeling**: 미래 상태를 통해 환경의 규칙을 스스로 학습
  - **Self-Reflection**: 에이전트가 자신의 행동에 대한 결과를 보고, 전문가의 행동이나 목표 상태와 비교함

→ 이를 통해 모델 스스로 어떤 행동이 잘못됐는지 분석(self-supervision)



Introduction

- Early Experience 장점
  - 보상을 주지 않아도 학습 가능
  - 사람이 직접 라벨링하지 않아도 가능
  - 에이전트가 스스로 개선 가능(환경 상태가 변화하면 이러한 경험이 학습 데이터가 됨)
- Contribution
  - Early Experience 개념 제안: RL과 SFT를 합친 새로운 개념으로, 보상이 없어도 경험을 학습하는 방법
  - Implicit world modeling과 self-reflection 학습 방법 제안: 경험을 신호로 변환하는 방법
  - 대규모 실험을 통한 Early Experience의 성능 향상 입증



Related Work

- World Model: RL 연구에서 상태를 보고, 다음 상태(보상)를 예측하는 모델
  - World model을 이용해 환경을 직접 시뮬레이터처럼 대체하는 방식
  - 최근의 world model은 LLM 자체가 해당 역할을 수행하도록 함(다음 상태에 대해 텍스트 기반 예측)
  - 기존 연구에서는 world model을 별도의 시스템처럼 따로 두고, 에이전트가 시뮬레이터를 사용하는 방식

→ 본 논문에서는 world model을 따로 구축하지 않고, 별도의 시뮬레이터도 필요 X

Related Work

- 기존의 self-reflection은 LLM에게 LLM의 응답에 대한 이유를 질답하는 방식 → 추론 능력 향상
  - 이러한 reflection은 프롬프트 안에서만 이루어지고, 학습하는 형태는 X (가중치 업데이트 X)
- 이후 다른 연구에서 reflection을 학습 신호(보상)로 활용하려는 시도가 진행됨
  - 하지만 RL 환경에서 보상이 있는 경우에만, reflection을 학습 데이터로 사용할 수 있었음

→ 본 논문에서는 에이전트가 자신의 행동을 보고, 정답이 틀렸다면 틀린 부분에 대해 스스로 분석하여 reflection을 다시 학습 데이터로 사용

Preliminaries

- LLM 에이전트 문제는 Markov Decision Process(MDP)와 같이 구조화 가능
- MDP: $M = (S, A, T, R, \gamma, \rho_0)$
  - S: 가능한 모든 상태
  - A: 가능한 모든 행동
  - $T(s, a)$: 환경 변화
  - $R(s, a)$: 보상
  - γ: Discount factor
  - ρ_0: 초기 상태 분포

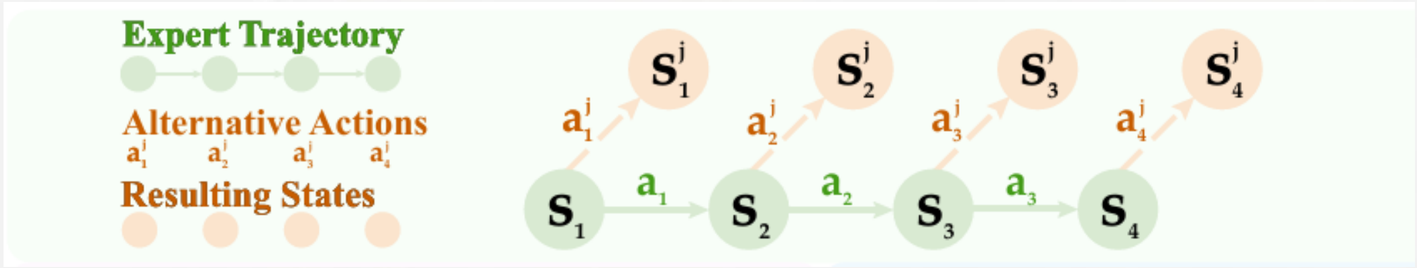
Preliminaries

- SFT Loss

$$L_{IL}(\theta) = -\sum_{i=1}^{N} \log \pi_\theta(a_i|s_i)$$

  - a_i: 전문가의 행동
  - s_i: 상태
  - 에이전트의 정책 $\pi_\theta(a_i|s_i)$가 전문가의 행동과 같아지도록 학습
  - SFT는 단순히 전문가의 행동을 모방하는 imitation learning

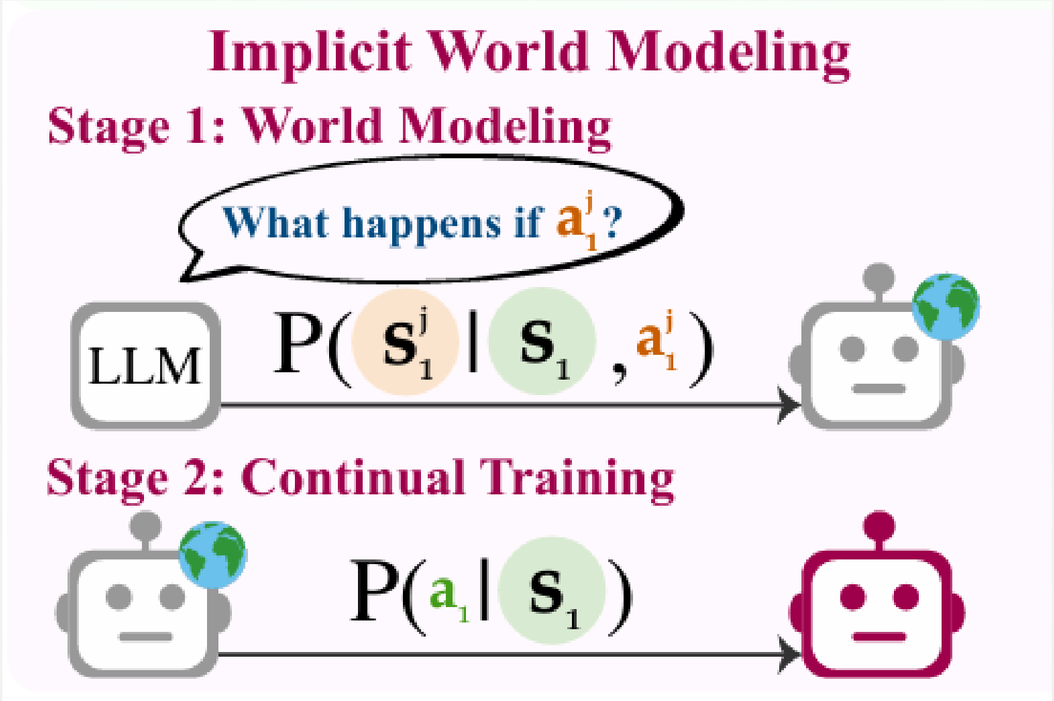
Methodology



- 원본 데이터 D_{expert}
  - $D_{expert} = \{(s_i, a_i)\}_{i=1}^{N}$

- 대체 행동 Candidate action set A_i
  - $A_i = \{a_i^1, a_i^2, \cdots, a_i^K\}$
  - 1개의 전문가 행동과 K개의 대체 행동 → 총 K+1개의 행동 후보 고려
- 각 행동을 실제로 실행해서 future state(s_i^j) 수집
- 미래 상태는 보상 없이 행동의 결과를 그대로 반영하고, 이후에도 변화 없이 원 상태 변화 그대로 사용

- 최종 데이터셋 $D_{rollout}$
  - $D_{rollout} = \{(s_i, a_i^j, s_i^j)\} | \ i \in [N], j \in [K]\}$



- **Implicit World Modeling(IWM)**
  - 대체 행동을 실제 환경에서 실행해서 얻은 다음 상태를 예측 시, 다음 토큰 예측으로 예측하게 만들면 모델이 world modeling 가능 (next token prediction fine-tuning)
- World Modeling Loss

$$L_{IWM} = -\sum_{(s_i, a_i^j, s_i^j) \in D_{rollout}} \log p_\theta \left(s_i^j \mid s_i, a_i^j \right)$$

  - 상태 변화 자체를 라벨로 두고, 모델이 현재 상태에서 행동을 했을 때 다음 상태가 나옴을 그대로 예측하도록 학습
- World model과 policy model을 따로 만들지 않고, 하나의 LLM이 두 개의 예측을 모두 수행

- **Self-Reflection(SR)**
  - 에이전트가 실제로 해본 대체 행동과 그 결과를 보고, 결과에 대한 추론을 학습 데이터로 사용

- Reflection 데이터 D_{refl}
  - $D_{refl} = \{(s_i, a_i^j, c_i^j)\}$
  - c_i^j : 결과에 대한 추론 (자연어 설명)

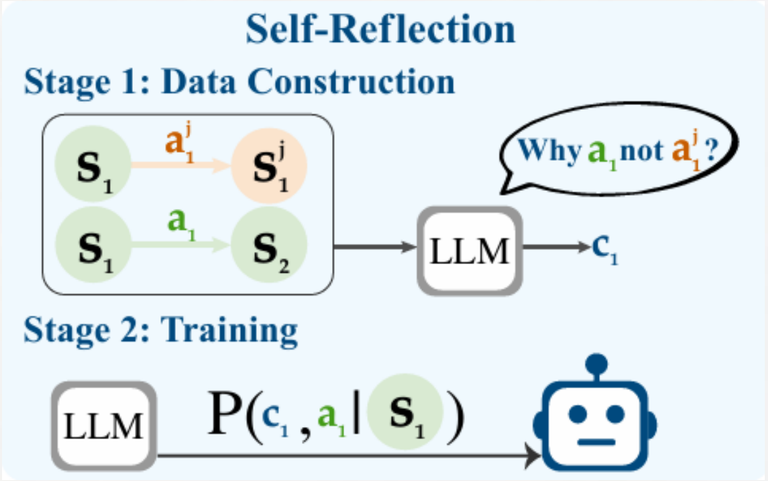
- Self-reflection Loss

$$L_{SR} = -\sum_{(s_i, a_i^j, c_i^j) \in D_{refl}} \log p_\theta(c_i^j \mid s_i)$$

  - 상태 변화 자체를 라벨로 두고, 모델이 현재 상태에서 행동을 했을 때 다음 상태가 나옴을 그대로 예측하도록 학습

- 현 상태에서 전문가처럼 판단할 수 있도록 reflection 결과를 정답 토큰으로 보고 예측

→ 모델이 판단(reasoning) 기준을 학습 가능함



Self-Reflection Prompt Template

You will be presented with a situation where you need to choose between multiple possible actions. Your task is to analyze the situation and provide reasoning about why we decide to take the expert action.

- **Situation Description (s_i):** {Situation Description}
- **Expert Action (a_i):** {Expert Action}
- **Expected Outcome (s_{i+1}):** {Future State of Expert Action}
- **Alternative Actions:**
  1. Action a_i^1: {Alt Action 1}, resulting state s_i^1: {State 1}
  2. Action a_i^2: {Alt Action 2}, resulting state s_i^2: {State 2}
  3. ...

Provide a detailed self-reflection as an *internal monologue* that demonstrates your reasoning process for the current situation. Your monologue should:

1. Analyze the situation and the goal.
2. Compare the possible actions, explaining why each may be less optimal.
3. Justify why the expert action is most suitable, grounded in the expected outcome.
4. Highlight any relevant clues, constraints, or consequences from the situation.

Guidelines:

- Stay strictly within the provided information.
- Avoid meta-commentary about being an AI.
- Use natural, step-by-step reasoning.
- Focus on logical decision-making.

Output: Directly write the self-reflection monologue, no extra headings, disclaimers, or external notes.

Experiments

- Model
  - Llama 3.2 - 3B
  - Qwen-2.5 - 7B
  - Llama 3.1 - 8B

- Evaluation
  - Success Rate(%)
  - F1-score: SearchQA만 적용(답변 품질 필요)

- Training Strategy
  - 학습 시, 동일한 prompt formatting 적용
  - IWM은 첫번째 에폭에만 IWM loss 적용, 이후 IL과 같은 loss로 진행

Environment	Description	# Traj.	# $\mathcal{D}_{\text{expert}}$
	MISC (Embodied and Scientific Simulation, and Travel Planning)		
ALFWorld (Shridhar et al., 2021)	Embodied instruction-following tasks in a simulated household, combining textual descriptions with high-level symbolic actions. We follow the setting of Feng et al. (2025).	3,553	21,031
ScienceWorld (Wang et al., 2022)	An interactive science lab simulator rendered in natural language, where agents perform multi-step experiments using tools and materials. We implement the gym (Brockman et al., 2016) for this environment.	1,000	14,506
TravelPlanner (Xie et al., 2024a)	Long-horizon travel planning tasks that require generating and refining multi-day itineraries using various tools and databases. We focus on the sole-planning mode and implement the gym for such an environment.	45	1,395
	Multi-Turn Tool-Use		
BFCLv3 (Patil et al., 2025)	Multi-turn tool-use tasks from the Berkeley Function Call Leaderboard v3, where agents interact with a Python-based API environment that simulates functional programs. We focus on the multi-turn tool use.	125	1,264
Tau-Bench (Yao et al., 2025)	Realistic customer-service scenarios requiring agents to interact with LM-simulated users, perform multi-turn tool use via APIs, and adhere to domain-specific policy documents. We focus on the Retail subset.	452	5,239
SearchQA (Jin et al., 2025)	Multi-hop question answering in open-domain settings, where agents issue search queries and reason over retrieved snippets to answer complex questions. We follow Search-R1 (Jin et al., 2025) settings and treat Musique as the in-domain dataset and HotpotQA, 2WikiMultiHopQA, and Bamboogle as out-of-domain datasets.	2,082	7,691
	Web Navigation		
WebShop (Yao et al., 2022)	Shopping tasks in a simulated e-commerce site, where agents must navigate, filter, and select the correct product based on natural language queries. We follow the setting of Feng et al. (2025).	1,571	15,464
WebArena-Lite (Zhou et al., 2024) (Liu et al., 2025)	Web navigation tasks across domains like e-commerce, forums, and content management. We follow Koh et al. (2024) to evaluate results with accessibility tree as observation space.	554	7,044

Experiments

- Effectiveness
  - 거의 모든 환경에서 Early Experience가 크게 향상
  - IWM: 구조화된 dynamics 있는 환경에서 강함
  - SR: reasoning 필요한 long-horizon task에서 강함
  - SFT보다 Early Experience 방식이 성능 향상시킴

Benchmark	Model	Prompt	Imitation Learning	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	20.3	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**)

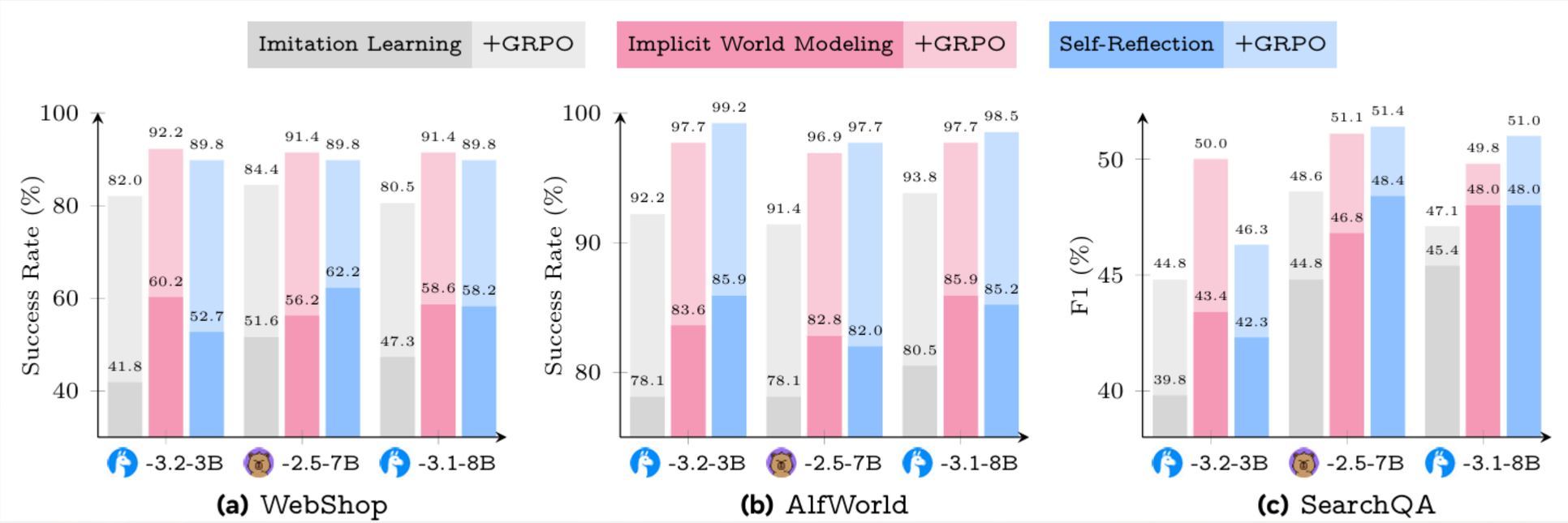
Experiments

- Out-of-domain generalization
  - Early Experience로 훈련된 모델은 reward가 없는 환경에서도 환경 변화를 읽고 행동을 조정하는 능력이 생겨, 새로운 상황에서도 더 높은 성공률을 보임

	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)

Experiments

- Compatibility with RL
  - RL을 바로 하면 sample efficiency가 낮고 reward sparse 문제 발생
  - Early Experience로 초기화하면 RL이 훨씬 빠르고 안정적으로 성능 향상 보임



Experiments

- Comparison to Baselines
  - Long CoT는 쉬운 환경에서만 약간 향상, 어려운 환경에서는 효과 사라짐
  - STaR-style은 hallucination이 많고, 전문가의 행동과 불일치가 많아서 오히려 성능 떨어짐
  - Early Experience(IWM, SR)를 두 baseline과 비교했을 때, 성능 향상 입증

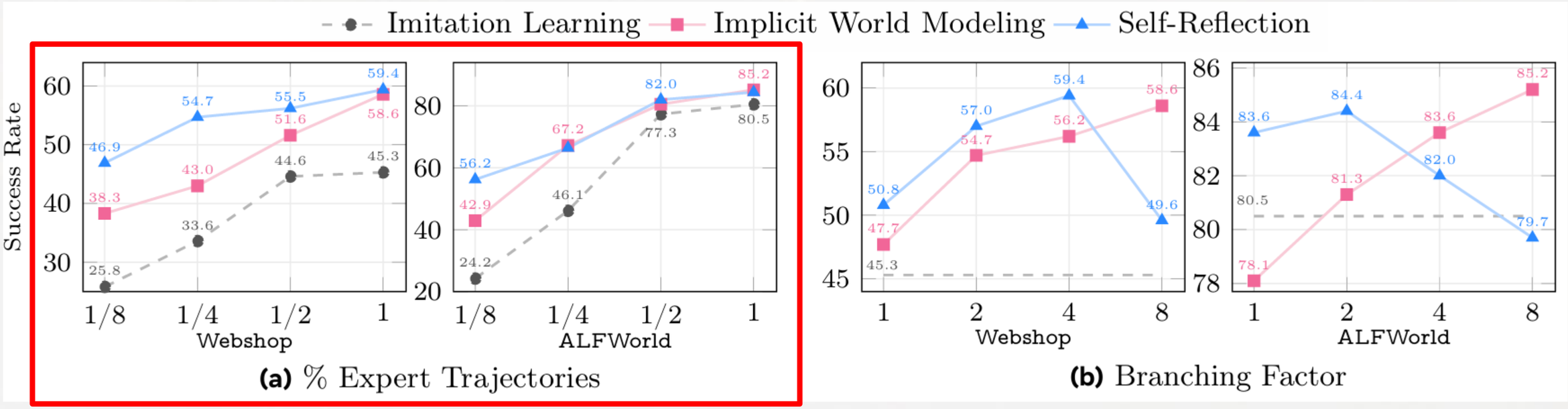
Table 4 Comparison of early experience with three representative baselines. All results are based on Llama-3.1-8B-Instruct.

	WebShop	ALFWorld
Prompt	0.0	25.0
+Long CoT	1.6 (+1.6)	28.4 (+3.4)
Imitation Learning	47.3	80.5
+Long CoT	0.0 (-47.3)	25.8 (-54.7)
+STaR	25.0 (-22.3)	74.2 (-6.3)
Ours-IWM	58.6 (+11.3)	85.9 (+5.4)
Ours-SR	58.2 (+10.9)	85.2 (+4.7)

- STaR-style data
  모델이 expert 행동에 대해 이론적 해석을 생성
  → 전문가의 행동과 일치하는 경우만 사용

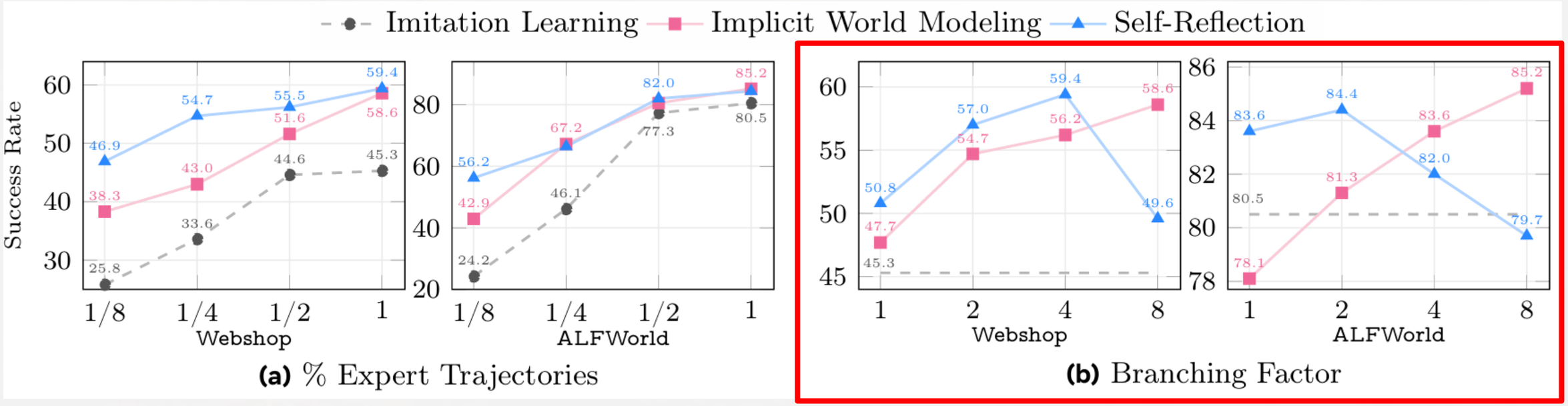
Experiments

- Impact of Amount of Human Data
  - Early Experience는 아주 적은 expert data만으로도 imitation learning 전체를 뛰어넘음
  - WebShop에서 1/8 데이터로 전체 Imitation Learning을 추월함
  - ALFWorld에서 ½ 데이터로 추월함



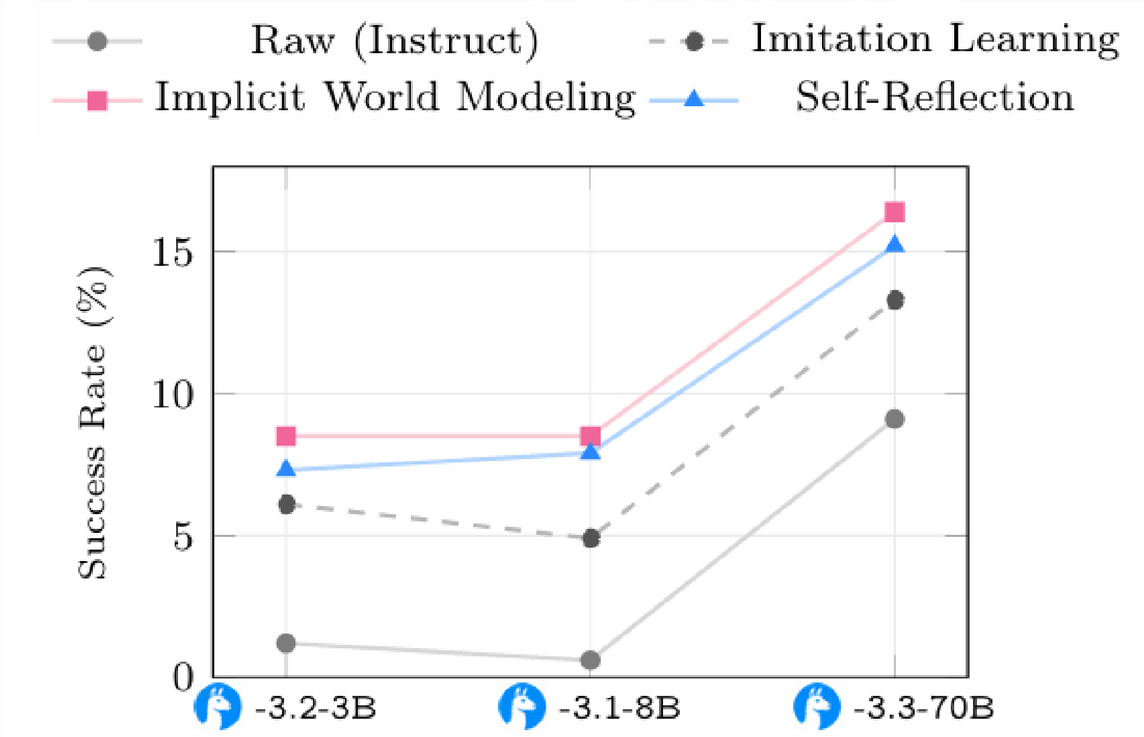
Experiments

- Impact of Branching Factor
  - Branching Factor(K, 각 상태에서 생성하는 대안 행동 수) 변화 실험
  - IWM은 K가 커질수록 성능 향상됨: 환경 전이 패턴을 더 많이 학습하기 때문
  - SR은 작은 K(2~4)에서 가장 좋은 성능을 보이며, 너무 큰 K에서는 오히려 노이즈가 증가하는 경향 있음



Experiments

- Model Scaling
  - 모델 크기와 무관하게 Early Experience가 계속 이득을 줌
  - 특히 WebArena-Lite 같은 어려운 환경에서 초대형 모델도 early experience 적용 시 성능 상승
  - LoRA fine-tuning만으로도 IWM/SR은 강력한 개선 제공



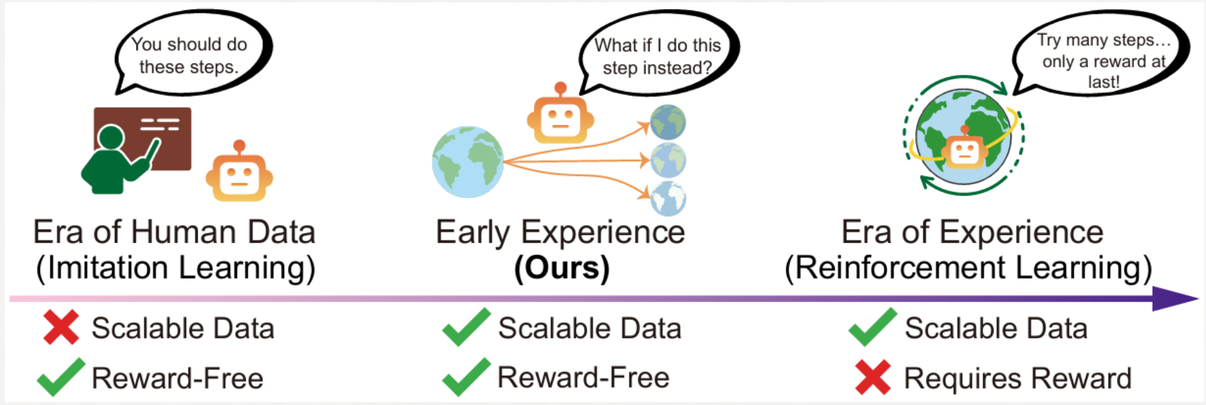
[WebArena-Lite benchmark 결과]

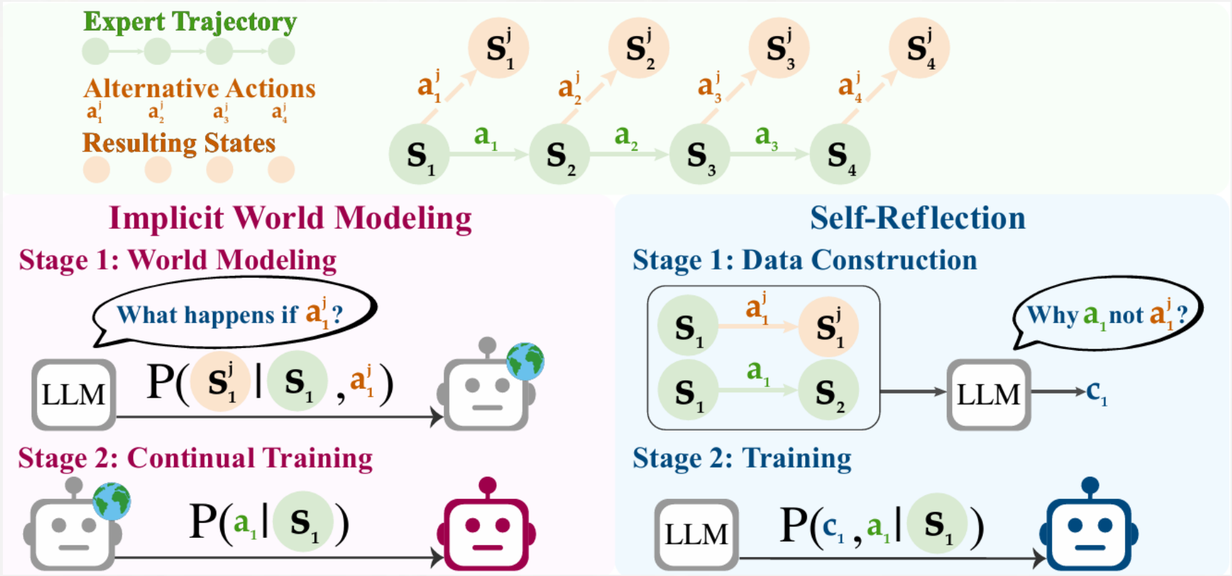
Conclusion

- Early Experience: 모델이 스스로 여러 대안 행동을 실행하고, 그 결과로 얻은 상태 변화(next state)를 보상 없이 감독 신호(supervision)로 사용

- Limitation
  - Short-horizon 중심
    - IWM, SR 모두 짧은 시퀀스(short-horizon) 중심으로 설계됨
    - 장기 의존성(long-horizon credit assignment)을 명시적 보상 없이 해결하는 것은 여전히 어려움
  - Implicitness의 한계
    - IWM은 명시적 world model이 아니라 암묵적 패턴 학습 → 복잡한 장기 추론에는 부족할 수 있음
  - 결과 신호가 전이(next state)에만 기반
    - 환경 반응이 제한적일 때 성능이 제한될 수 있음

Agent Learning via Early Experience





- Motivation
  - LLM 기반 에이전트는 RL 환경이 부족하고, 보상 제공이 어렵기 때문에 전이 규칙이나 행동 결과를 학습할 기회가 매우 제한적임
  - 이를 해결하기 위해, 모델이 스스로 여러 대안 행동을 실행하고 그 결과 상태를 감독 신호로 사용하는 reward-free 학습 패러다임(Early Experience)을 제안
- Methodology: Early Experience
  - Implicit World Modeling(IWM): (state, action)으로 다음 상태를 예측하게 만들어, 에이전트가 환경 전이 규칙을 암묵적으로 학습하도록 함
  - Self-Reflection(SR): 대안 행동들의 결과 상태를 비교하게 하고, 왜 특정 행동이 더 나은지 자연어로 스스로 설명(CoT)하게 만듦