Agent Skill

2026.08.27

김명진

Introduction: Skill 외재화 연구의 공통 골격

Skill Library 계열이 공유하는 5단계 파이프라인

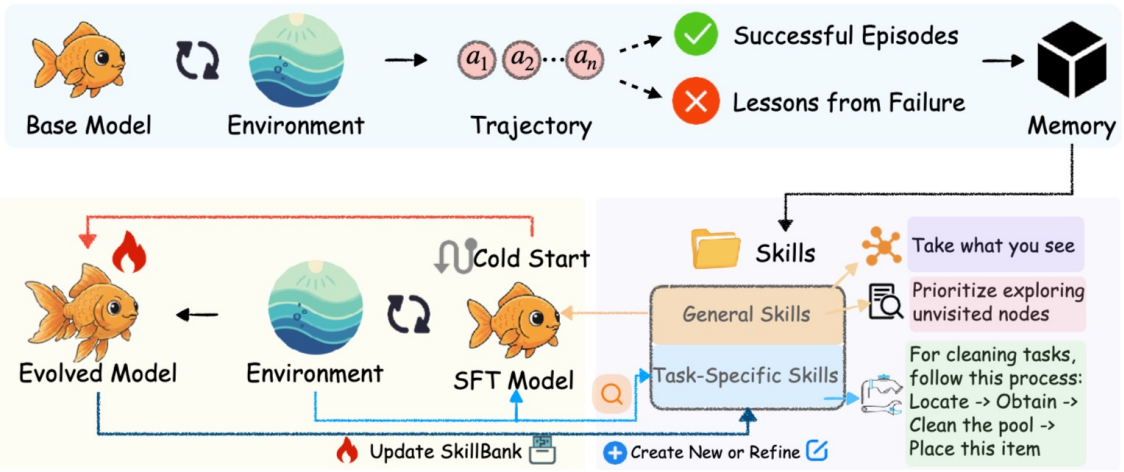
- 경험 수집 — agent가 환경에서 rollout을 수행하고, 성공한 episode와 실패에서 얻은 교훈을 함께 모음
- Skill 추출·추상화 — 개별 trajectory를 다른 task에도 적용 가능한 절차·규칙 형태로 일반화
- SkillBank 축적 — general skill(탐색 원칙)과 task-specific skill(도메인 절차)로 계층화해 외부에 저장
- Retrieval & 주입 — 새로운 task가 들어오면 관련 skill을 검색해 prompt·context에 넣어줌
- RL로 정책 갱신 — skill을 잘 활용하도록, 또 좋은 skill을 만들도록 보상을 설계해 정책을 학습

- → 세부 구현은 달라도, skill이 모델 바깥의 저장소에 존재하고 추론 시점에 불러온다는 점은 모두 동일

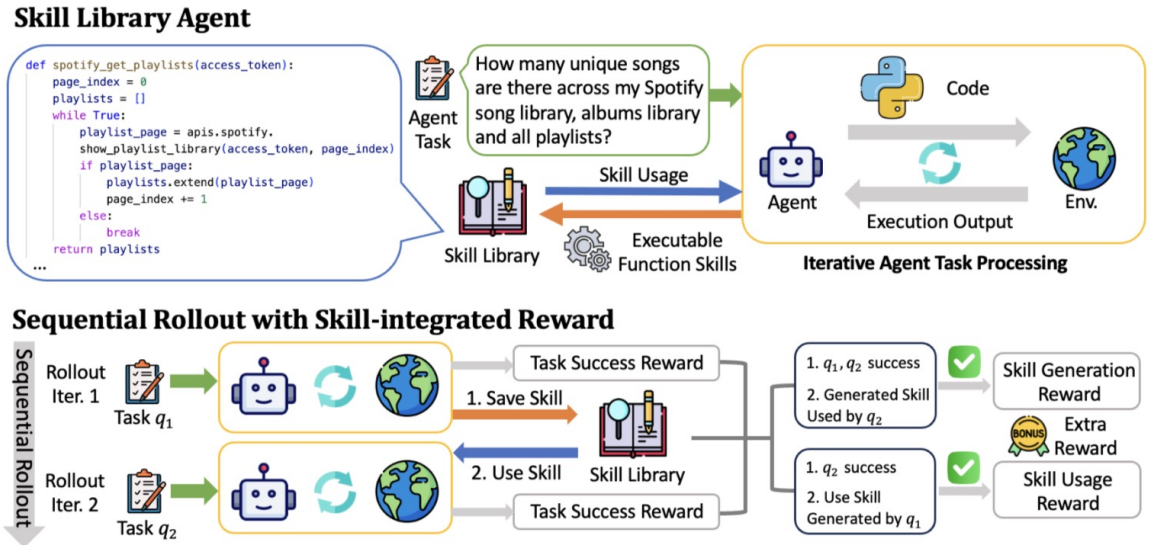
Introduction: 대표 연구 둘러보기

SkillRL & SAGE

- 두 연구 모두 같은 골격 위에 있지만, skill의 형태와 보상 설계가 다름



SkillRL (https://arxiv.org/abs/2602.08234)



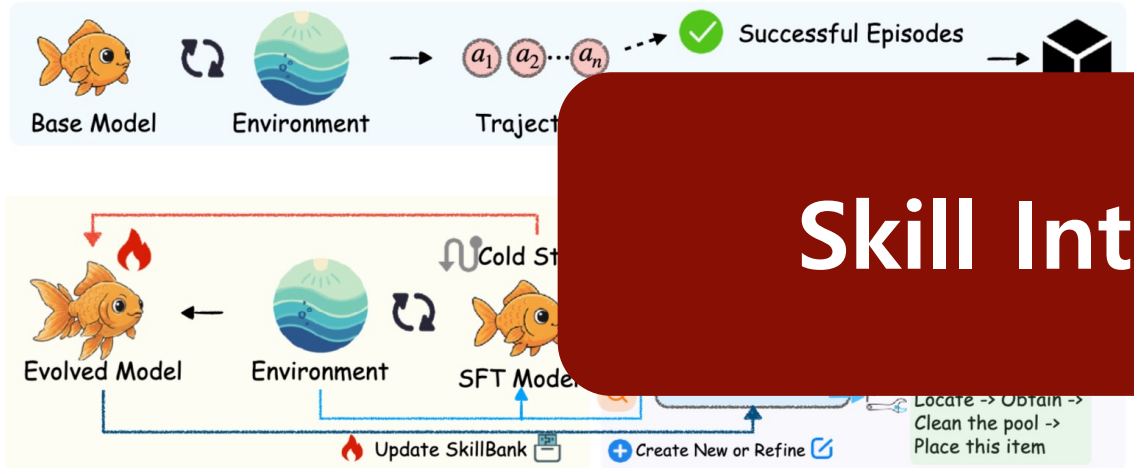
SAGE (https://aclanthology.org/2026.acl-long.69/)

- SkillRL — 자연어 skill을 general / task-specific 2계층 SkillBank로 관리, policy와 함께 co-evolve
- SAGE (ACL 2026) — skill이 실행 가능한 python 함수, Sequential Rollout으로 skill 재사용 여부를 확인
- SAGE는 task success reward에 skill generation·usage reward를 더한 skill-integrated reward 설계

Introduction: Skill Internalization으로의 전환

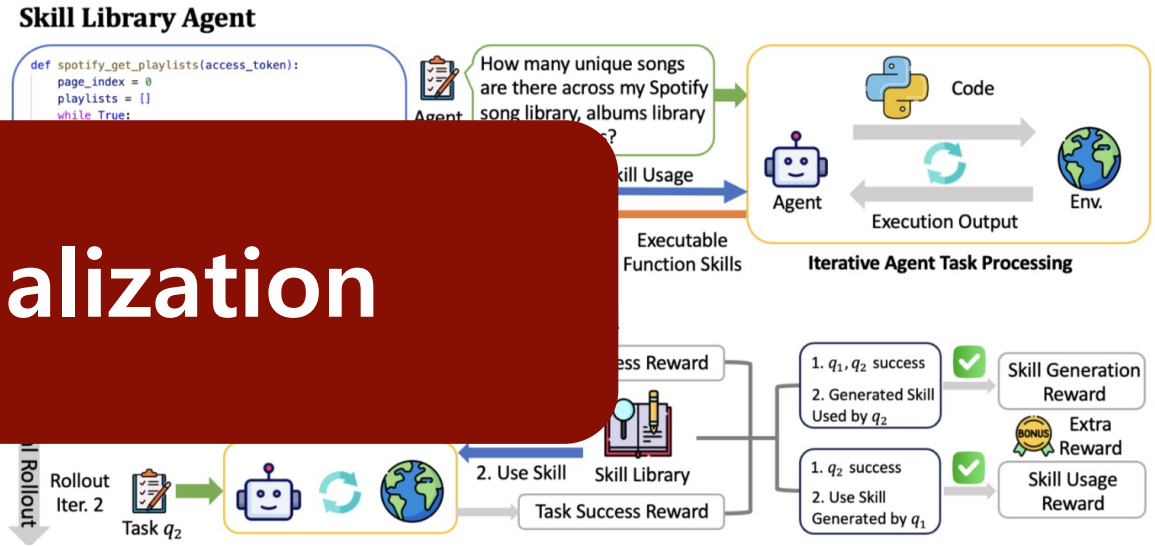
학습이 끝나면 skill은 어디에 있어야 할까?

- 그렇다면 skill을 계속 '들고 다니는' 대신, 학습 과정에서 모델 파라미터로 흡수시킬 수는 없을까?



SkillRL (https://arxiv.org/abs/2602.08234)

Skill Internalization



SAGE (https://aclanthology.org/2026.acl-long.69/)

- Skill Internalization — 학습 때만 skill을 쓰고 추론 시에는 skill 없이 동작하는 정책을 만드는 문제 설정
- 추론 시 retrieval·저장소·teacher가 전혀 필요 없고, 성능은 skill을 쓰는 방법과 대등하거나 그 이상이어야 함
- 오늘 다룰 두 논문: Skill0 (문제를 처음 정식화, curriculum 기반) / OVCSD (outcome-verified self-distillation, 최신)

LongCat

SKILL0: In-Context Agentic Reinforcement Learning for Skill Internalization

Zhengxi Lu[1,2⋆], **Zhiyuan Yao**[2], **Jinyang Wu**[3], **Chengcheng Han**[2], **Qi Gu**[2†]
Xunliang Cai[2], **Weiming Lu**[1], **Jun Xiao**[1], **Yueting Zhuang**[1], **Yongliang Shen**[1†]
[1]Zhejiang University [2]Meituan [3]Tsinghua University
{zhengxilu, syl}@zju.edu.cn guqi03@meituan.com

arXiv 2026.05.15

Motivation & Contribution

Motivation

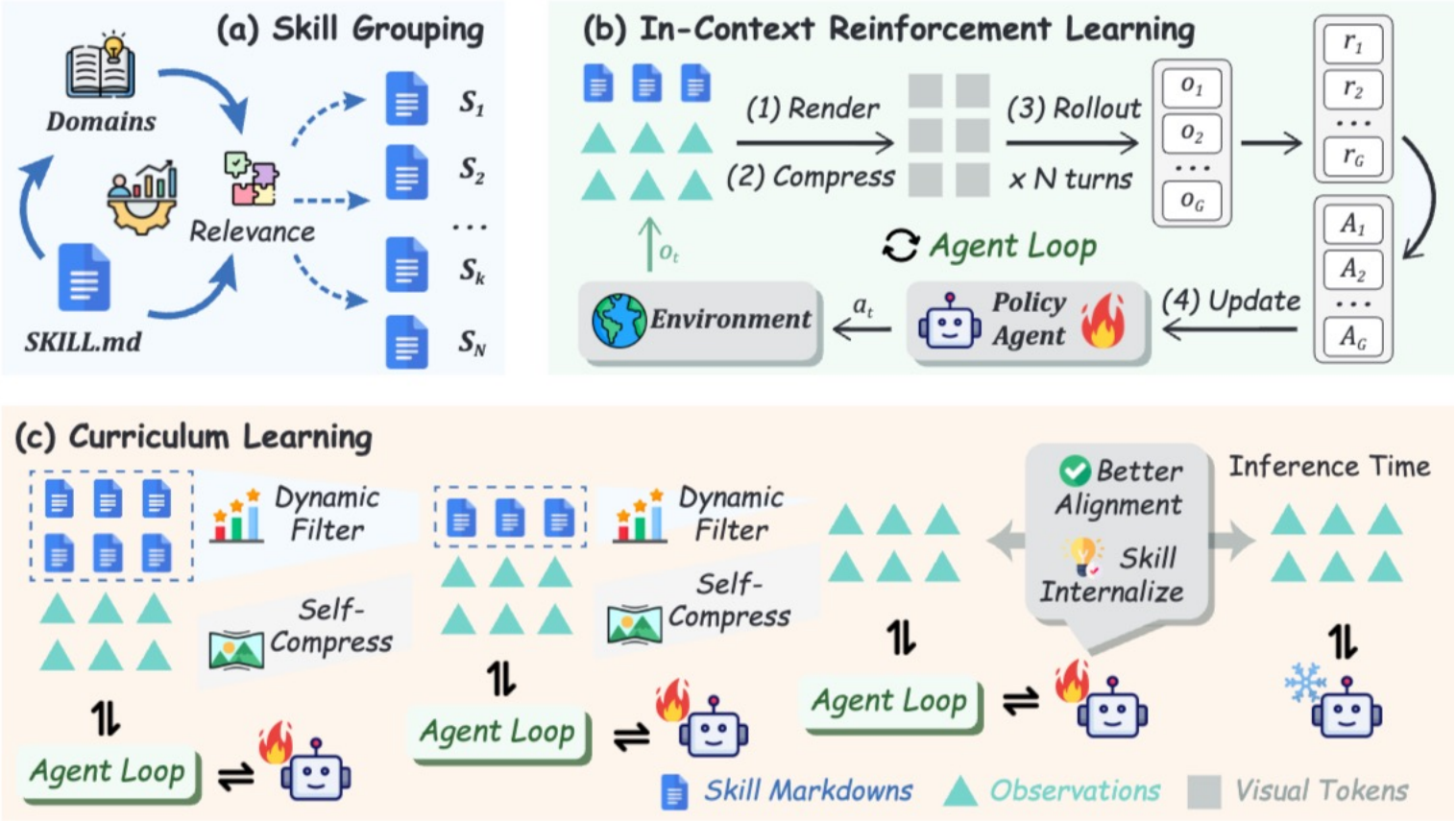
- Agent Skill — 절차적 지식과 실행 자원을 담은 구조화된 패키지를 추론 시점에 동적으로 RAG하는 방식이 LLM agent 능력 확장의 표준 메커니즘으로 자리잡음

- 그러나 inference-time skill augmentation은 세 가지 근본적 한계를 지님
  1. Retrieval noise — 무관하거나 오도하는 guidance가 agent의 context를 오염시킴
  2. Token overhead — 주입된 skill 내용이 multi-turn 상호작용마다 누적되어 확장성을 제한
  3. 실제 학습이 아님 — prompt의 skill을 따르는 모델은 skill을 실행할 뿐, 능력은 context에 존재

- 핵심 질문: skill을 모델 파라미터에 internalize하여, 추론 시 retrieval 자체를 불필요하게 만들 수 있는가?

Methodology: Skill0

Overview



Methodology: Skill0

Agent Loop

- 목표: context를 텍스트가 아닌 이미지로 렌더링, 압축하여 토큰 비용을 줄이며 multi-turn 상호작용 수행
- Task 정의 — 매 step t에서 관측 o_t를 받아 $\pi_\theta(a_t \mid I, h_t)$로 행동을 샘플링, 환경은 $o_{t+1} = E(o_t, a_t)$로 전이
  history $h_t = \{o_1, \ldots, o_t\}$, 과제 완료 또는 최대 step 도달까지 반복
- 1. Render — history h_t와 활성 skill 집합 S를 monospace로 그린 compact RGB 이미지로 변환
  ALFWorld·WebShop 10pt / 최대 392px, Search-QA 12pt / 최대 560px — instruction 검정, observation 파랑, action 빨강으로 의미별 색상 구분
- 2. Compress — vision encoder가 이미지를 visual token으로 인코딩: $V_t = Enc(h_t, S\,;\, c_t)$
  압축률 $c_t \in (0,1]$을 정책이 행동과 함께 스스로 생성: $(a_t, c_t) \sim \pi_\theta(a_t, c_t \mid I, V_t)$
- 3. Rollout ×N turns → 4. Update — group 단위 보상으로 정책 갱신 (다음 장)
- 전체 history를 유지하면서도 step당 0.5k 미만의 토큰만 사용

Methodology: Skill0

In-Context Reinforcement Learning (ICRL)

- 목표: 학습 rollout에서 skill을 주입하고, 그 결과를 GRPO로 정책 파라미터에 반영

- Reward — 과제 성공 보상 r_t에 압축 보상 r_t^{comp}를 결합
  $r_t^{comp} = \ln(c_t)$ if $I_{succ}(\tau) = 1$, 아니면 0 → $\tilde{r}_t = r_t + \lambda \cdot r_t^{comp}$
  성공한 trajectory에만 압축 보상을 부여해 과제를 포기하고 압축만 하는 붕괴를 방지, log 형태는 압축의 한계효용 체감을 반영

- Advantage — 같은 query q에 대해 G개 trajectory를 샘플링한 뒤 group 내에서 정규화
  $A_i = (\tilde{r}(\tau_i) - \mu_G) / (\sigma_G)$, μ_G, σ_G는 group 내 보상의 평균과 표준편차

- Objective — token-level importance ratio에 clipping + reference policy에 대한 KL 정규화

$$\mathcal{L}_{\text{SKILL0}}(\theta) = \mathbb{E}_{\tau_i \sim \pi_{\theta_{\text{old}}}(q),\, q \sim \mathcal{D}} \frac{1}{\sum_{i=1}^{G} |\tau_i|} \sum_{i=1}^{G} \sum_{t=1}^{|\tau_i|} \text{clip}(r_{i,t}(\theta),\, A_i,\, \epsilon) - \beta \cdot \mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]$$

Methodology: Skill0

Adaptive Curriculum Learning

- 목표: skill을 '언제' 회수할지를 고정 스케줄이 아니라 정책의 현재 상태에 맞춰 결정
- (a) Offline skill grouping — 학습 전 validation set을 N개 sub-task T_k로 분할
  각 T_k는 skill 파일 S_k와 도메인, 목표가 대응되도록 구성 → skill마다 전용 검증 sub-task 확보
- (b) Online helpfulness-driven curriculum — d step마다 각 skill의 helpfulness Δ_k를 측정
  $\Delta_k = Acc(\pi_\theta, T_k, S) - Acc(\pi_\theta, T_k, \emptyset)$ — skill을 넣었을 때와 뺐을 때의 성능 차이
  정책이 내재화할수록 $Acc(\emptyset) \rightarrow Acc(S)$, 즉 $\Delta_k \rightarrow 0$이 되어 자동으로 회수 대상이 됨
- Skill budget — N_s개 stage에 걸쳐 선형 감소: $M^{(s)} = \lfloor N \cdot (N_s - s) / (N_s - 1) \rfloor$
  stage당 약 $N/(N_s - 1)$개씩만 줄어 context 분포 변화가 완만 → 급격한 policy collapse 방지
- 매 stage 절차 — Filter ($\Delta_k > 0$만 유지) → Rank (Δ_k 내림차순) → Select (상위 $m \leq M^{(s)}$개)
  $M^{(s)} = 0$이 되면 $S = \emptyset$ — 이후 학습과 추론 모두 완전 skill-free로 진행

Experiments

Evaluation task and metrics

- Evaluation task

  1. ALF-World

     텍스트 기반 embodied 환경 / 6개 task type (Pick, Look, Clean, Heat, Cool, Pick2)

     Input: 지시문 + 관측 텍스트 → Output: 이동·조작 액션 시퀀스 / Metric: Success Rate (%)

  2. Search-based QA

     In-domain: NQ, HotpotQA / Out-of-domain: TriviaQA, PopQA, 2Wiki, MuSiQue, Bamboogle (E5 retriever)

     Input: 질문 → Output: search 질의 반복 후 최종 답변 / Metric: Accuracy (%)

  3. Webshop

     실제 쇼핑몰을 모사한 웹 환경, 고정된 128개 validation task

     Input: 제품 요구사항 문장 → Output: search[...], click[...] 액션 시퀀스 / Metric: Score, Accuracy (%)

- Models & Setup

  Qwen2.5-VL-3B / 7B-Instruct, 4×H800에서 최대 180 step 학습 ($N_s = 3$, $d = 10$, $G = 8$)

  Baselines: Zero/Few-Shot, GRPO, AgentOCR, EvolveR, SkillRL — †는 추론 시 skill 사용, *는 시각 인코딩

Results

Table 1: **Performance on ALFWorld and Search-QA tasks.** We report the success rate (%) and the average context token cost (k) per step. † denotes models validated with skill augmentation; ⋆ denotes methods that encodes visual context with reduced token overhead. We simply reproduce results of SkillRL-3B without cold start and skill evolution. **Best** and second-best are highlighted.

	ALFWorld								Search-QA								
Method	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**	**Avg↑**	**Cost↓**	**NQ**	**Triv**	**Pop**	**Hotp**	**2Wk**	**MuS**	**Bam**	**Avg↑**	**Cost↓**
Qwen2.5-(VL)-3B-Instruct																	
Zero-Shot	27.0	24.3	4.5	20.5	10.2	0.0	15.2	1.21	9.4	31.3	19.8	15.0	14.8	4.7	16.8	15.9	0.48
Few-Shot†	44.1	45.1	30.5	44.1	9.2	3.3	29.3	2.30	11.8	30.9	20.2	13.7	18.4	4.5	25.6	17.9	0.86
Zero-Shot⋆	44.3	27.6	8.6	3.1	5.7	3.1	17.6	0.48	10.2	27.7	10.9	9.1	12.2	3.7	15.2	12.7	**0.15**
Few-Shot⋆†	57.1	25.3	4.5	5.5	10.2	9.4	23.8	0.88	11.1	26.2	14.2	15.4	13.4	3.0	19.2	14.6	0.36
GRPO	92.6	85.7	70.6	86.6	**79.3**	65.0	79.9	1.02	39.3	**60.6**	41.1	**37.4**	34.6	**15.4**	26.4	36.4	0.61
AgentOCR⋆	91.9	81.8	76.0	73.3	76.1	70.0	78.2	0.38	38.6	56.5	41.7	33.6	30.7	14.6	24.0	34.2	0.26
EvolveR	77.3	24.5	47.9	41.7	24.6	22.5	44.1	1.89	**43.4**	58.4	**43.4**	37.3	**38.1**	13.7	32.8	38.2	–
SkillRL†	91.9	**100**	82.9	**87.4**	78.7	70.0	82.4	2.21	38.6	57.6	40.3	33.6	31.1	13.3	58.1	38.9	0.87
SKILL0⋆	**95.6**	80.4	**100**	86.7	78.7	**75.2**	**87.9**	**0.38**	39.8	57.5	42.3	35.1	33.7	13.3	**63.7**	**40.8**	0.18
Qwen2.5-(VL)-7B-Instruct																	
Zero-Shot	67.6	35.4	19.3	31.3	30.1	4.4	31.3	1.08	10.4	32.4	22.3	15.8	15.4	7.2	19.2	17.5	0.70
Few-Shot†	75.4	64.9	67.5	26.7	19.4	8.9	48.4	2.12	12.3	36.8	24.5	17.7	18.2	6.5	24.8	20.1	0.97
Zero-Shot⋆	46.0	35.6	19.1	7.1	5.5	5.4	21.1	0.52	6.9	30.4	12.0	10.5	9.1	5.5	24.0	14.0	**0.26**
Few-Shot⋆†	44.3	55.4	52.9	0.0	11.2	5.4	28.9	1.79	10.5	31.9	18.7	14.2	14.4	6.9	24.8	17.3	0.41
GRPO	92.6	93.8	85.2	80.0	82.7	56.5	81.8	0.95	45.1	**63.7**	44.0	**43.6**	**43.2**	16.8	37.6	41.9	0.73
AgentOCR⋆	95.6	**96.2**	78.1	73.2	72.4	72.0	81.2	0.43	43.1	61.0	45.4	40.8	38.3	15.7	36.8	40.1	0.36
EvolveR	64.9	33.3	46.4	13.3	33.3	33.3	43.8	–	43.5	63.4	**45.9**	38.2	42.0	15.6	54.4	43.1	–
SkillRL†	97.9	71.4	90.0	**90.0**	**95.5**	**87.5**	**89.9**	–	**45.9**	63.3	**45.9**	43.2	40.3	**20.2**	**73.8**	**47.1**	–
SKILL0⋆	**100**	85.8	**94.6**	81.9	85.7	80.1	89.8	**0.41**	42.7	61.1	45.3	40.0	38.3	16.4	66.9	44.4	0.34

- Skill0(*)는 추론 시 skill 없이도 ALFWorld Avg 87.9(3B) / 89.8(7B) 달성 — AgentOCR 대비 +9.7%p
- context 비용도 최소: ALFWorld 0.38k vs SkillRL† 2.21k tokens/step, Search-QA 0.18k vs 0.87k

Results

Table 2: **Performance on WebShop (128 tasks).** We report the Score (%), Acc (%) and the average context token cost (k) per step. All methods encode visual context with reduced token overhead.

Method	Inference w/ Skills	Qwen2.5-VL-3B Score↑	Qwen2.5-VL-3B Accuracy↑	Qwen2.5-VL-3B Tokens↓	Qwen2.5-VL-7B Score↑	Qwen2.5-VL-7B Accuracy↑	Qwen2.5-VL-7B Tokens↓
Baselines							
Zero-Shot		23.4	6.3	0.46k	26.8	7.8	0.48k
Few-Shot	✓	45.8 [+22.4]	18.4 [+12.1]	0.78k [+0.52k]	51.2 [+24.4]	24.2 [+16.4]	0.91k [+0.43k]
AgentOCR		75.2 [+51.8]	56.3 [+50.0]	0.42k [-0.04k]	78.6 [+51.8]	59.3 [+51.5]	0.40k [-0.08k]
Ours: RL w/ Skills							
SKILL0	✓	80.9 [+57.5]	64.1 [+57.8]	0.94k [+0.48k]	85.3 [+58.5]	71.9 [+64.1]	0.89k [+0.41k]
SKILL0		78.6 [+55.2]	66.4 [+60.1]	0.49k [+0.03k]	85.1 [+58.3]	74.2 [+66.4]	0.46k [-0.02k]

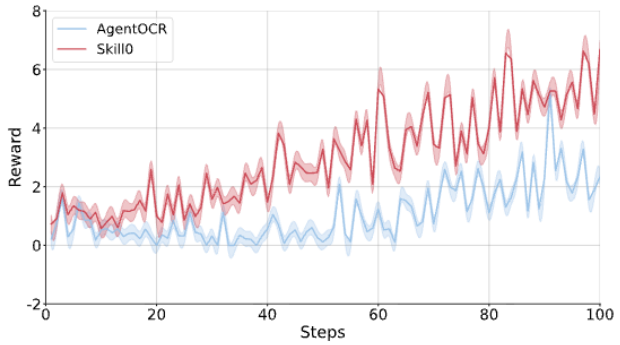


Figure 3: Comparison of training dynamics with AgentOCR on Qwen2.5-VL-3B.

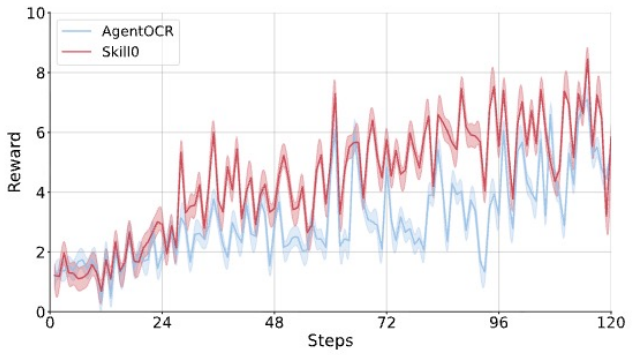


Figure 4: Comparison of training dynamics with AgentOCR on Qwen2.5-VL-7B.

- WebShop 128 tasks — 3B Score 78.6 / Acc 66.4, 7B Score 85.1 / Acc 74.2
- 추론 시 skill을 다시 넣어도 성능이 거의 같음 → 이미 파라미터에 내재화
- Zero-Shot 대비 Acc +60.1%p(3B), +66.4%p(7B) / 토큰은 오히려 −0.02k
- Fig 3·4 — 3B·7B 모두 AgentOCR보다 높은 reward를 유지하며 상승

Results

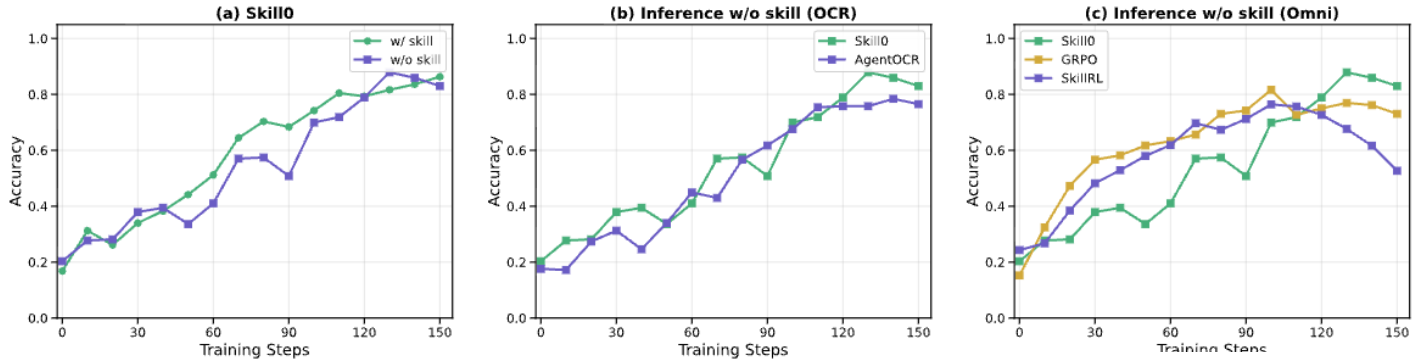


Figure 5: **Training Dynamics Comparison.** (a) Validation perfor without skill augmentation, evaluated every 10 training steps. (b) SKILL0 and AgentOCR, both evaluated without skill augmentatio SKILL0 (OCR) against GRPO (Text) and SkillRL (Text), all evalu

- (a) skill 유무별 validation 성능 — 초반엔 skill이 있을 때 빠르게 오르지만 후반에는 skill 없는 쪽이 역전
- (b) 같은 skill-free 조건에서도 Skill0 > AgentOCR → 향상이 skill 설명 의존이 아니라 내재화된 지식에서 옴
- (c) skill-free 조건에서 GRPO·SkillRL은 조기에 정체되지만 Skill0는 꾸준히 상승
- Fig 6 — helpfulness Δ_k의 rise-then-fall: 초반 낮음 → 중반 상승(skill에 grounding) → curriculum이 skill을 걷어내며 0으로 수렴

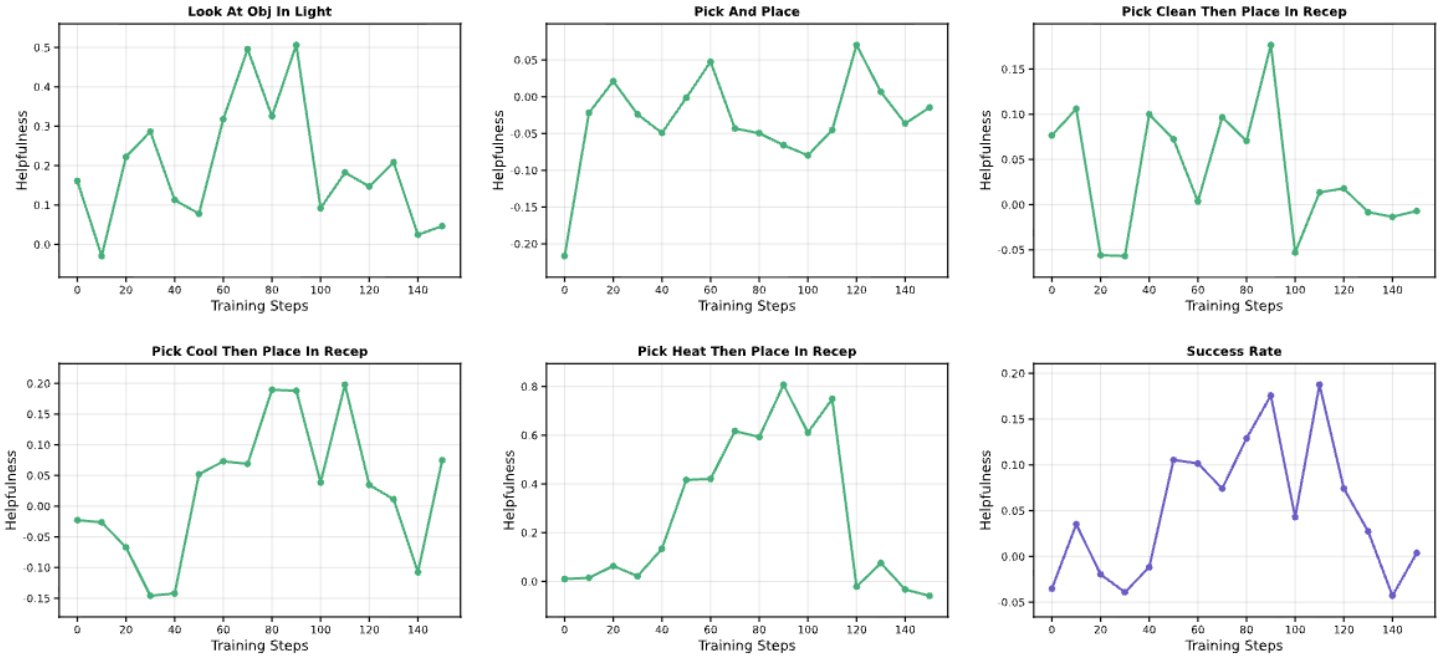


Figure 6: **Training Dynamics of *Helpfulness*,** which are reported by Δ_k for each sub-task k.

Ablations

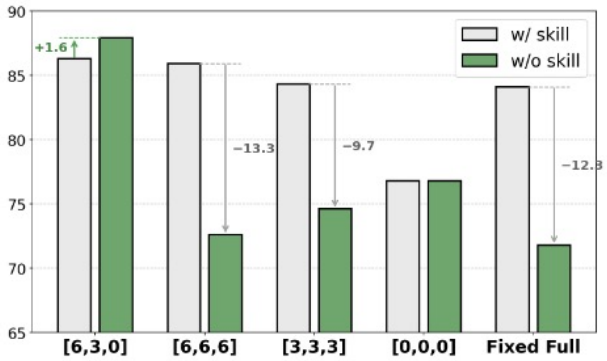


Figure 7: Ablations of skill budget M.

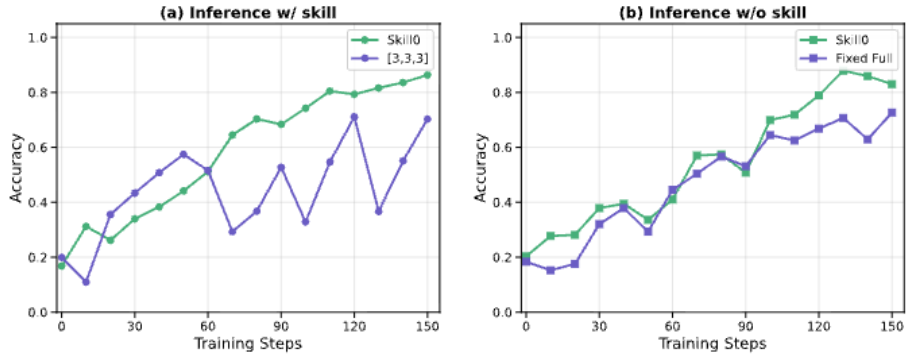


Figure 8: Ablations of skill budget during training process.

Table 3: Ablations of Dynamic Skill Curriculum on ALFWorld in different inference settings.

Method	w/ $\mathcal{S}$	w/o $\mathcal{S}$	Δ
`Filter & Rank & Select`	86.3	87.9	↑1.6
w/o `Filter`	81.6	78.9	↓2.7
w/o `Rank` (Random Select)	76.6	62.9	↓13.7

Table 4: Impact of Validation Interval d on ALFWorld and Search-QA (subset).

d	ALFWorld	Search-QA
10	87.9	48.9
5	87.5	49.6
20	78.1	42.3

- Fig 7 — skill budget M: [6,3,0]만 skill 제거 시 오히려 +1.6%p (내재화 성공)
- 고정 예산 [6,6,6] / [3,3,3] / Fixed Full은 −13.3 / −9.7 / −12.3%p로 붕괴
- Fig 8 — [3,3,3]은 초기 탐색 제한으로 불안정, Fixed Full은 skill-free 추론에서 지속 열세
- Table 3 — Filter 제거 −2.7%p, Rank 제거(무작위 선택) −13.7%p
- Table 4 — 검증 주기 d = 10이 최적, d = 20은 78.1로 −9.8%p

Conclusion

- Skill0 — agent skill을 추론 시점의 외부 자원이 아니라 학습으로 파라미터에 흡수시킬 대상으로 재정의, skill internalization을 명시적 학습 목표로 정식화한 최초의 RL framework
- In-Context RL (학습 때만 skill 주입) + Dynamic Curriculum (helpfulness Δ_k 기반 Filter – Rank – Select) + 시각 렌더링 기반 self-compression의 결합
- ALFWorld +9.7%p, Search-QA +6.6%p, WebShop Acc +10.1%p를 step당 0.5k 미만의 토큰으로 달성
- skill을 제거했을 때 오히려 성능이 오르는 positive transfer (Δ = +1.6%p) 확인 — 내재화가 실제로 일어났음을 보여주는 직접 증거

From Scoring to Acting: Outcome-Verified Comparative Self-Distillation for LLM Agents

Xu Xia[1,2], **Jinghua Piao**[2,3]*, **Min Yang**[2,4], **Xiaochong Lan**[3], **Jiaju Chen**[2,5], **Yong Li**[2,3]*

[1]Southeast University [2]Zhongguancun Academy [3]Tsinghua University
[4]Shandong University [5]University of Science and Technology of China
s-xx25g@bza.edu.cn {pjh22, lanxc22}@mail.tsinghua.edu.cn minyang@mail.sdu.edu.cn cjj01@mail.ustc.edu.cn
ysgong@sdu.edu.cn liyong07@tsinghua.edu.cn

arXiv 2026.07.30

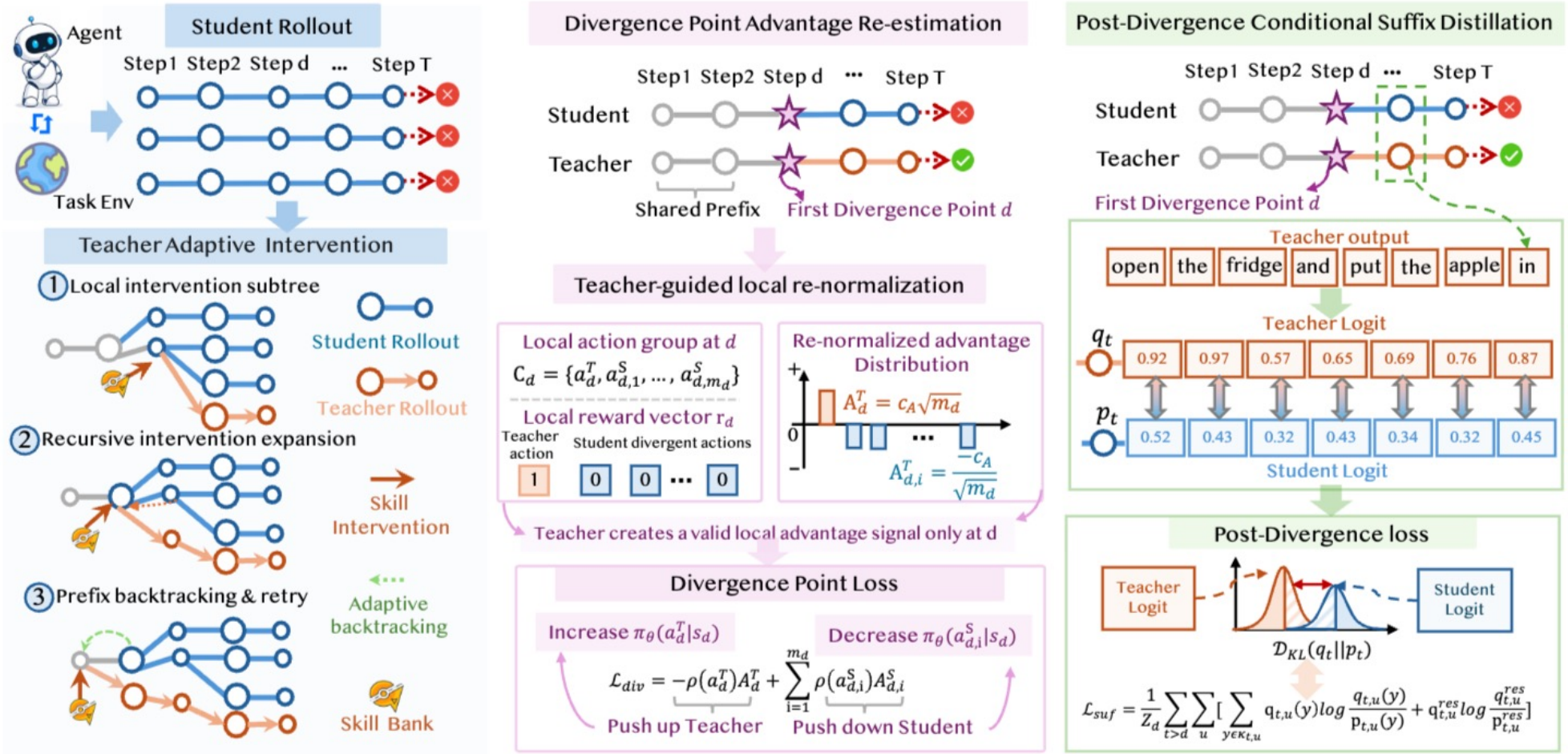
Motivation & Contribution

Motivation

- 최근 LLM agent 연구의 흐름 — external capability elicitation에서 capability internalization으로 이동
- 외부 정보(skill, memory, demonstration)로 얻는 이득은 대개 일시적 — 외부 정보가 사라지면 행동도 사라짐
- 기존 On-Policy Self-Distillation(OPSD)의 한계 — teacher가 student trajectory 위의 action을 채점(scoring)
  teacher의 국소적 선호 = 최종 성공 기여도라는 가정에 의존하지만, long-horizon task에서 이 가정은 특히 취약
  teacher가 환경과 상호작용하지 않으므로 그 선호는 task outcome으로 검증되지 않음
- 충분히 활용되지 못한 세 가지 정보원
  1. 실패한 student rollout — 구조적 정보를 담고 있음에도 그대로 폐기됨
  2. teacher trajectory를 균일하게 distill — 교정(correction)과 완수(completion) 행동이 구분되지 않음
  3. 두 정책의 행동이 갈리는 상태(divergence)에서 정렬·직접 비교되는 경우가 거의 없음

Methodology: OVCSD (Outcome-Verified Comparative Self-Distillation)

Overview



Methodology: OVCSD

Base Student Rollout

- 목표: 전부 실패한 rollout group에서 재사용 가능한 공유 state를 찾아내기
- 매 update마다 frozen skill-free policy가 K개 trajectory 생성: $D_x^0 = \{\tau_i^0 \sim \pi_{\bar{\theta}}^0\}_{i=1}^K$
  trajectory $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T)$, 종단 보상 $R(\tau)$
  GRPO baseline advantage: $\hat{A}_i = (R_i - \bar{R}) / (s_R + \varepsilon_{std})$
- 개입 조건 — 모든 trajectory가 실패하고 보상 분산이 사실상 0일 때만 발동
  $z_x^{fail} = I[\max_i R_i < r_{succ}] \cdot I[s_R \leq \varepsilon_R]$ — 이때 GRPO는 학습 신호를 전혀 얻지 못함
- Observation 1 — 실패한 rollout에도 공유 상태 구조가 남아 있음
  ALFWorld task group의 47.0%, WebShop의 52.8%가 all-failure group
  ALFWorld는 all-failure group의 100%에서 공유 prefix, coverage 85.8% (WebShop은 15.2% / 4.9%)
- → 실패 trajectory를 prefix tree로 구성. 노드 v는 깊이 d_v의 공유 상태이며, 그 prefix를 따르는 trajectory 부분집합 M(v)를 지지 (상태는 snapshot 또는 replay로 복원)

Methodology: OVCSD

Adaptive Outcome-Verified Intervention

- Skill-conditioned teacher — 학습 대상 모델에 skill context만 추가로 준 정책
  $\pi_{\bar{\theta}}^{S}(a_t \mid h_t, S_x) = \pi_{\bar{\theta}}(a_t \mid h_t, c(S_x))$. $c(S_x)$는 skill 정보의 인코딩
- 1. Deepest-first selection — prefix tree에서 가장 깊은 유효 공유 노드 v부터 선택
  student가 이미 이룬 진전을 최대한 보존하기 위함
- 2. Adaptive teacher execution — 상태 v에서 teacher가 최대 N_T개의 continuation 실행
  $\tau_v^T \sim \pi_{\bar{\theta}}^{S}(\cdot \mid h_v, S_x)$
- 3. Outcome verification — $R(\tau_v^T) \geq r_{succ}$ 인 continuation만 채택
  실패 시 가장 가까운 조상 노드로 backtrack 후 재시도 → 개입 지점을 적응적으로 확장
  검증된 continuation 하나가 $M(v)$의 모든 실패 branch를 감독하며, 조상 노드에서 중복 감독하지 않음
- 핵심 — teacher의 scoring이 아니라 환경이 검증한 결과만 supervision으로 사용
  teacher continuation의 43.5%가 검증에서 탈락 / 개입된 group의 74.3%가 최종적으로 검증된 branch 획득

Methodology: OVCSD

Alignment-Aware Comparative Learning

- 목표: teacher와 student의 행동이 처음 갈리는 지점에서 국소적으로 정렬하여 직접 비교

- First divergence point — 공유 prefix 이후 처음으로 행동이 달라지는 위치

$$d_{v,i} = d_v + \min\left\{ j : a^{\mathrm{T}}_{v,j} \neq a^{0}_{i,d_v+j},\ \ \mathrm{Align}\left(h^{\mathrm{T}}_{v,j'}, h^{0}_{i,d_v+j'}\right) = 1,\ \forall j' \leq j \right\}.$$

- Local comparison group — 같은 divergence 지점 (v, d)를 공유하는 $m_{v,d}$개의 실패 branch를 묶음

- Teacher-guided local re-normalization — 각 지점에서 양/음 advantage 총량을 맞춤

$$A^{T}_{v,d} = \sqrt{m_{v,d}}, \qquad A^{0}_{v,d,i} = -\frac{1}{\sqrt{m_{v,d}}}.$$

Methodology: OVCSD

Overall Objective and Inference

- Post-Divergence Suffix Distillation — 갈라진 이후 행동을 KL로 전수

  teacher 분포: $q_{t,m} = stopgrad[\, \pi_{\theta}^{S}(\cdot \,|\tilde{h}_{t,m}^{T}, S_x\,)\,]$

  student 분포: $p_{t,m} = \pi_{\theta}^{0}(\cdot \,|\tilde{h}_{t,m}^{T}\,)$ — student는 skill 없이 같은 history만 봄

$$\mathcal{L}_{\text{suf}} = \frac{1}{|\mathcal{D}^{+}|} \sum_{(v,d)\in\mathcal{D}^{+}} \frac{1}{|\Omega_{v,d}|} \sum_{(t,m)\in\Omega_{v,d}} D_{\text{KL}}(q_{t,m} \,\|\, p_{t,m})\,,$$

- Overall objective: $L_{OVCSD} = L_{GRPO}^{aug} + \lambda_{suf} \cdot L_{suf}$

  L_{GRPO}^{aug}는 divergence 지점의 국소 advantage를 반영, 나머지 행은 기존 advantage 유지

  divergence 지점 = 결정 교정(policy gradient), 이후 구간 = 완수 행동(distillation)으로 역할 분담

Experiments

Evaluation task and models

- Evaluation task

  1. ALF-World

     6개 task type (Pick, Look, Clean, Heat, Cool, Pick2), 종단 이진 보상

     Input: 가정 내 지시문 + 관측 텍스트 → Output: 이동·조작 액션 / Metric: Success Rate (%)

  2. Webshop

     다중 자연어 제약을 만족하는 상품을 검색·구매 / Input: 요구사항 문장 → Output: search[...], click[...]

     Metric: 공식 Score(×100)와 Success Rate(Acc)

- Models

  Qwen3-1.7B / Qwen2.5-3B / Qwen2.5-7B — 세 규모 모두에서 평가, 공통 SFT 초기화(Vanilla)에서 출발

  Baselines: Vanilla(SFT), SkillCOT*(추론 시 skill 검색), GRPO(skill-free RL), OPSD

  Skill-SD(teacher가 action을 채점), SDAR(초기 상태에서 skill-augmented rollout)

  SkillCOT*를 제외한 모든 방법은 skill-free deployment prompt 사용 → 동일 조건 비교

- Setup — group size K = 8, 100 update 학습 / ablation은 128-task validation set에서 50 update 기준

Results

- Main Results

Method	ALFWorld							WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	Avg	Score	Acc
Qwen3-1.7B									
Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	46.5	4.7
SkillCOT*	10.5	50.0	16.1	0.0	0.0	5.0	9.4	23.0	2.3
GRPO	71.1	41.7	36.4	40.0	31.8	31.6	46.1	67.3	38.3
OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	47.4	9.3
Skill-SD	52.9	37.5	69.2	42.9	60.0	36.8	52.3	**81.8**	53.9
SDAR	73.5	25.0	76.9	33.3	40.0	36.8	53.9	76.8	58.6
OVCSD	**96.9**	**83.3**	**91.7**	**75.0**	**71.4**	**74.8**	**83.6**	77.9	**62.5**
Qwen2.5-3B									
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	6.7	0.8
SkillCOT*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	0.2	0.8
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	79.8	63.3
OPSD	48.8	41.7	16.7	0.0	15.8	16.7	28.1	11.3	3.1
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	75.9	64.0
SDAR	**97.1**	62.5	**100.0**	61.9	75.0	84.2	84.4	85.0	68.0
OVCSD	89.6	**75.0**	89.6	**95.2**	**83.9**	**88.8**	**89.1**	**85.3**	**73.4**
Qwen2.5-7B									
Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	5.9	1.6
SkillCOT*	51.7	50.0	32.3	5.3	4.3	0.0	23.4	1.7	0.8
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	80.9	72.6
OPSD	50.0	60.0	22.7	21.4	17.6	9.5	32.8	4.5	2.3
Skill-SD	93.9	**93.8**	90.9	**100.0**	69.2	68.4	85.1	86.1	76.5
SDAR	94.7	75.0	**100.0**	86.7	68.2	78.9	85.9	89.4	82.8
OVCSD	**98.2**	83.3	97.9	**100.0**	**82.1**	**94.4**	**92.2**	**91.2**	**84.4**

Table 1: Performance on ALFWorld and WebShop tasks. We report success rate (%) on ALFWorld and score/accuracy (%) on WebShop. WebShop Score is reported as the task score multiplied by 100. * denotes evaluation with retrieved skills at inference time. **Bold with yellow box** and underlined with blue box mark the best and second-best results within each base-model block. All methods except Skill-COT* use the skill-free deployment prompt.

- Interaction Budget

Update	Student	Replay	Teacher	Overhead	Eff. K
30	74,084	152	1,818	2.59%	8.21
60	121,377	288	2,267	2.06%	8.17
100	176,961	492	3,314	2.11%	8.17

Table 2: Cumulative environment action steps by source on ALFWorld over 100 updates. *Overhead* is replay plus teacher continuations as a share of the total; *Eff.* K rescales the nominal group size of 8 by that total. Privileged interaction stays near 2% throughout training.

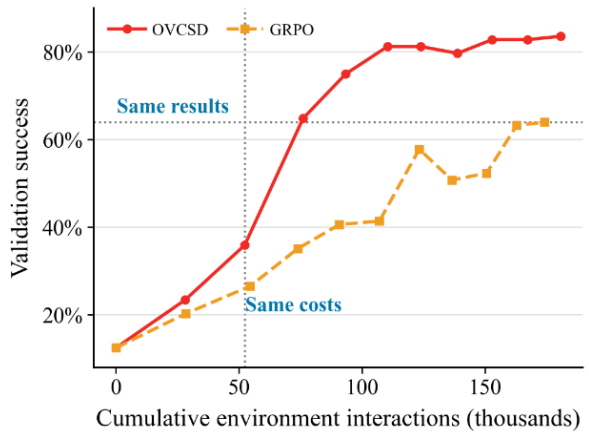


Figure 3: Skill-free validation success against cumulative environment interactions on ALFWorld (1.7B). At a matched budget of 123k actions OVCSD reaches 81.3% against GRPO's 57.8%. The shaded band marks budgets where only OVCSD has completed validation.

Ablations

How the branch is transferred

- 검증된 teacher branch를 student에게 어떻게 전달할 것인가 — 두 목적함수의 기여도 분해 (ALFWorld skill-free validation, 50 update)
- Local Â — divergence 상태에서의 국소 advantage (결정 교정)
- Suffix KL — 이후 완수 구간의 top-k distillation (완수 행동)
- 단독 +25.00 / +22.65%p, 결합 시 +39.84%p

Method	Local $\hat{A}$	Suffix KL	Success
GRPO			41.41
Local-PG only	✓		$66.41_{+25.00}$
Suffix-KL only		✓	$64.06_{+22.65}$
OVCSD	✓	✓	$81.25_{+39.84}$

Table 3: Objective ablation on skill-free ALFWorld validation (%). Local $\hat{A}$ is the localized advantage at the divergence state, Suffix KL the top-k distillation on the completion suffix. Subscripts are gains over GRPO.

How the branch is acquired

- 검증된 branch를 어떻게 획득할 것인가 — anchor 선택과 결과 검증의 효과
- Random anchor 79.69 (깊이 1.73) / No verification 80.47 (채택분의 43.9%가 실제로는 실패)
- OVCSD 81.25 — 깊이 3.56 유지, 71회 중 48회 검증, 검증 branch당 비용 25.60으로 최저

Strategy	Success	Attempts	Verified	Depth	Cost/ver.
Random anchor	79.69	40	25	1.73	33.00
No verification	80.47	66	37	2.74	30.57
OVCSD	**81.25**	71	48	3.56	**25.60**

Table 4: Intervention ablation over 50 updates. *Attempts* are teacher continuations executed, *Verified* those that complete the task; *Depth* is mean replay depth and *Cost/ver.* the replay plus teacher actions spent per verified branch. Attempt counts are an outcome of each run, not a matched input; the controlled comparison is cost per verified branch.

Ablations

What Intervention Changes

Scored action	GRPO	OVCSD	Δ	95% CI
Verified teacher	+0.060	+0.342	+0.282	[0.17, 0.41]
Failed student	−0.054	−0.081	−0.027	[−0.24, 0.17]
Prefix control	−0.208	−0.162	+0.046	[−0.13, 0.23]
States increased (%)	67.9	91.0	+23.1	[12.8, 34.6]

Table 6: Change in teacher-forced action-span log-probability relative to the shared SFT initialization.

- 학습이 실제로 '검증된 teacher 행동의 확률'을 올렸는가 — SFT 초기화 대비 teacher-forced action-span log-prob 변화
- 검증된 teacher 행동: GRPO +0.060 → OVCSD +0.342 (Δ +0.282)
- 실패한 student 행동·prefix control은 거의 변화 없음 → 국소적으로만 작동
- 확률이 오른 정렬 divergence 상태 비율: 67.9% → 91.0%

Case Study

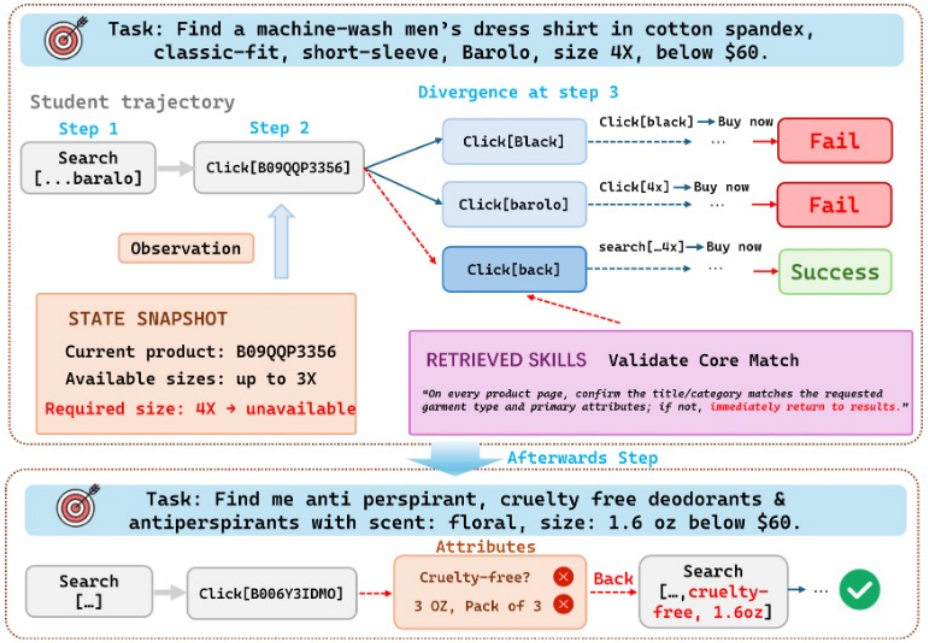


Figure 4: Case study.

- student는 사이즈가 없는 페이지에서 옵션만 반복 클릭하다 실패, teacher는 'Validate Core Match'로 검색 결과로 되돌아가 누락 제약을 추가

Conclusion

- OVCSD — teacher의 '점수(scoring)'가 아니라 환경이 검증한 '행동(acting)'을 supervision으로 사용하는 on-policy self-distillation

- 실패한 student rollout을 prefix tree로 조직
  → student가 도달한 상태에서 skill-conditioned teacher를 적응적으로 호출 → outcome-verified continuation만 채택

- divergence 지점에서는 state-aligned 국소 비교로 결정을 교정하고, 이후 구간은 top-k KL로 완수 행동을 전수

- ALFWorld·WebShop에서 최강 baseline 대비 각각 최대 +29.7 / +5.4%p, privileged interaction은 3% 미만, 동일 예산에서 GRPO 대비 2.1× sample efficiency

감사합니다