Tencent 腾讯

从交互执行到经验学习：Code Agent 的能力提升闭环探究

浙江大学博士生、腾讯Codebuddy & Workbuddy青云计划研究实习生

郝哲正

2026年7月

Code Agent 如何在训练-测试-部署的闭环中持续提升能力?

目录

训练层面 SFT vs RLVR 中的数据编排

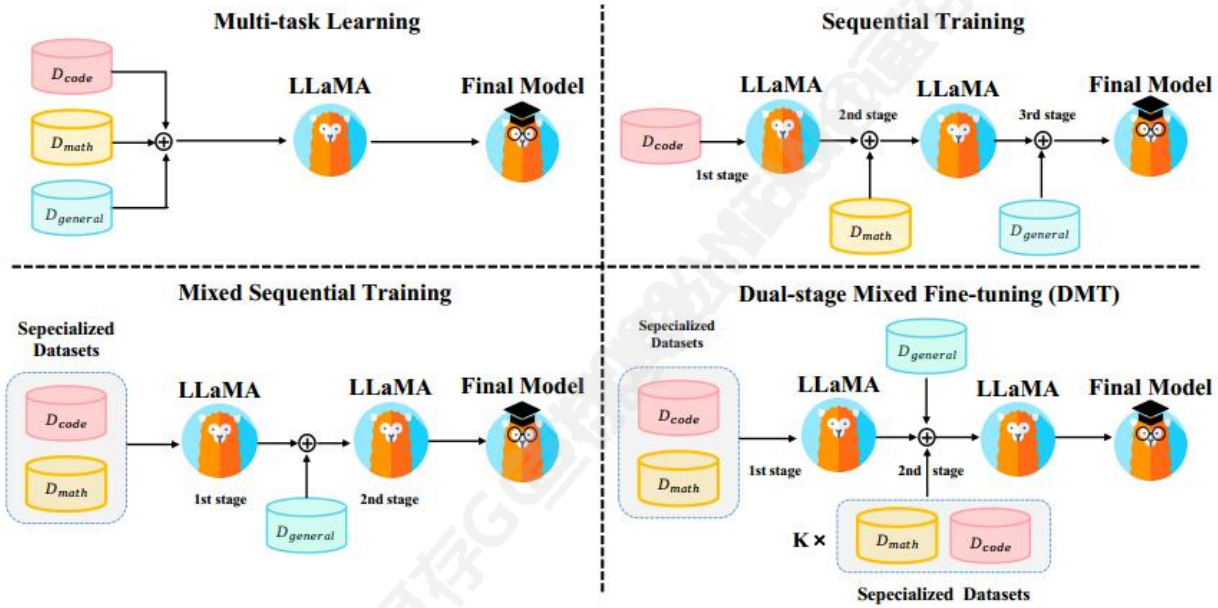
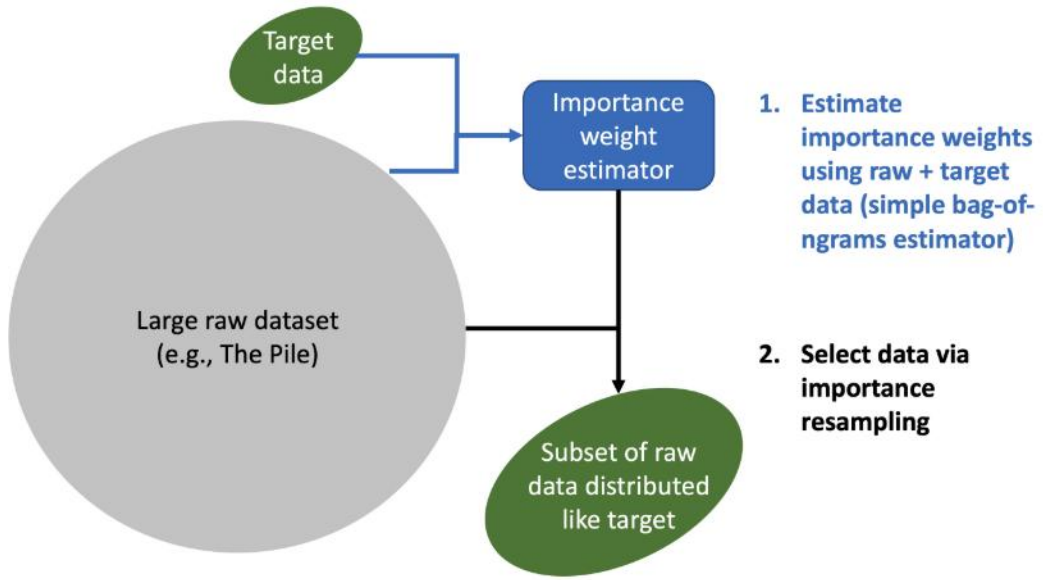


Figure 1: The illustration of four different training strategies in this paper.



Data Scheduling Formulation

$$\mathcal{J}_{\text{schedule}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\ldots}\left[\omega(q,t)\cdot(\text{original objective term for } q)\right]$$

ACC-based Scheduling

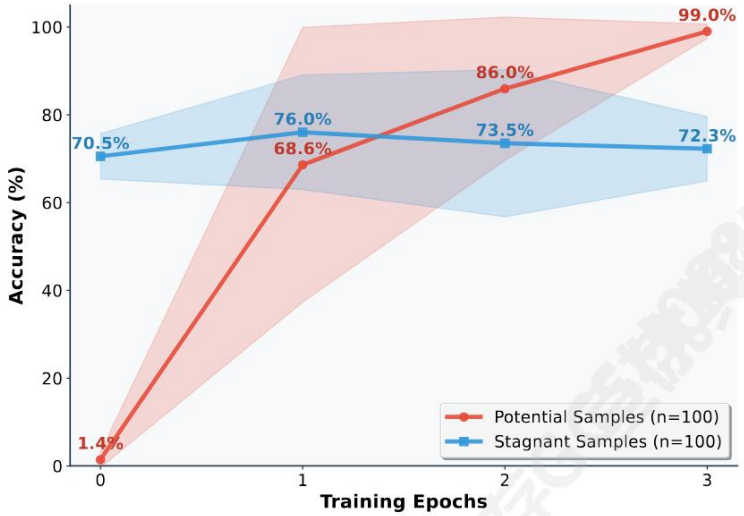
$$\alpha(\text{ACC}(q),t) = (1-\gamma(t))\text{ACC}(q) + \gamma(t)(1-\text{ACC}(q))$$

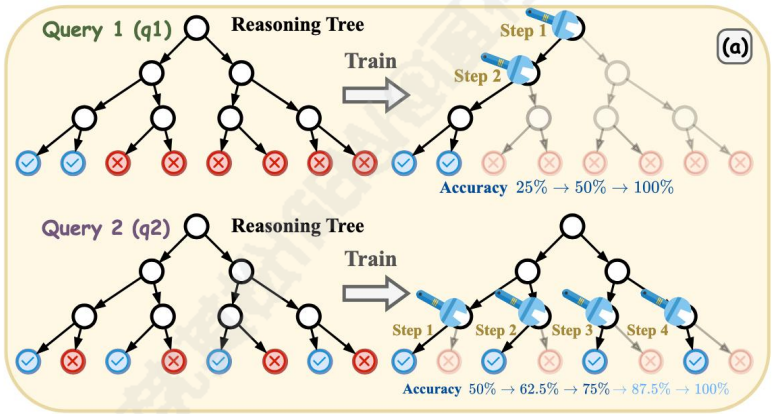
$$\omega = \text{rank}(\alpha)\%\cdot\omega_{\max} + (1-\text{rank}(\alpha)\%)\cdot\omega_{\min}.$$

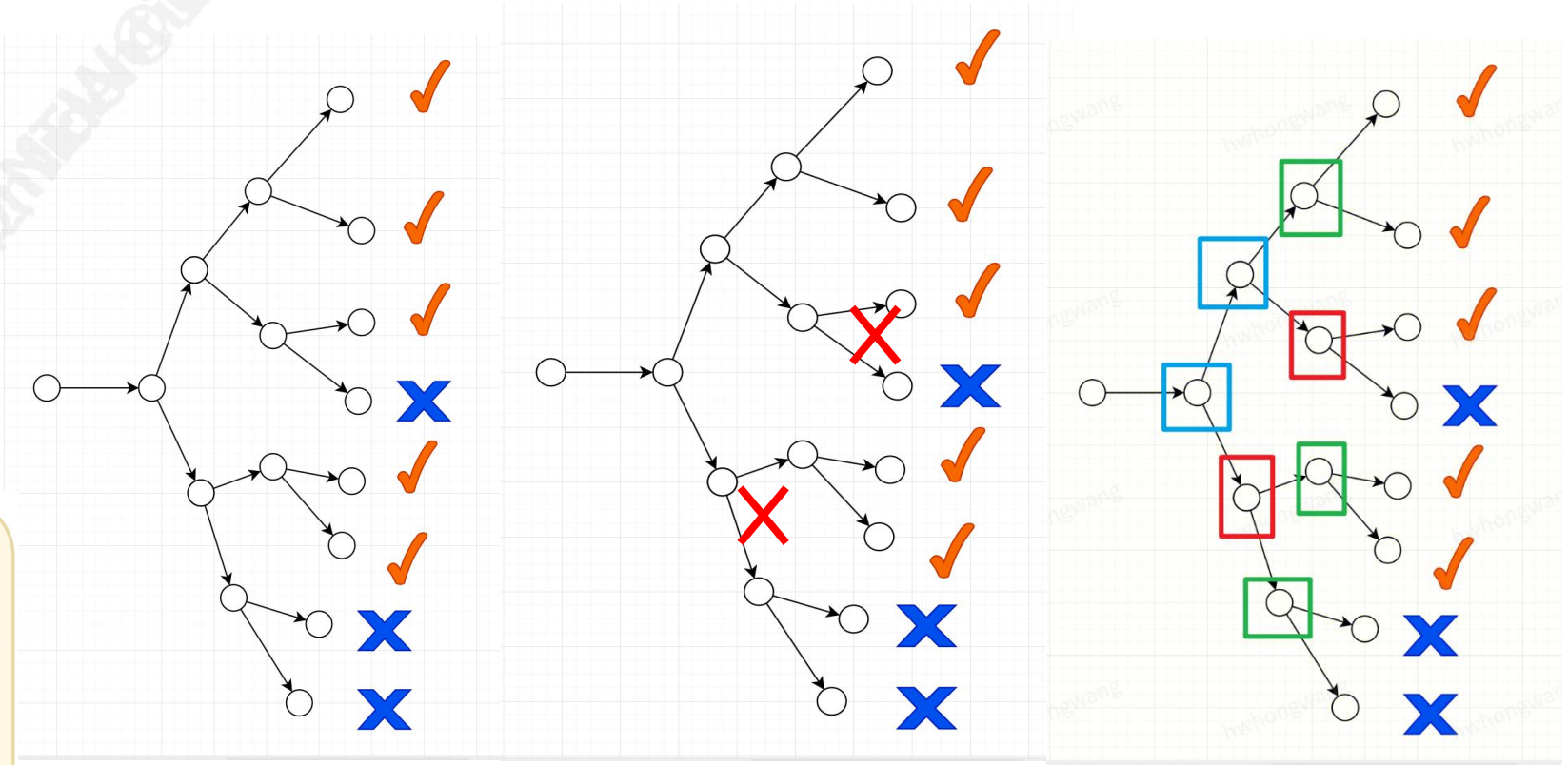
Wang H, Hao Z, Luo J, et al. Scheduling Your LLM Reinforcement Learning with Reasoning Trees. In ICLR, 2026.

Agentic RL 中的数据编排

关键问题：在Agentic RL中，什么样的度量可以衡量一个数据的训练难度？





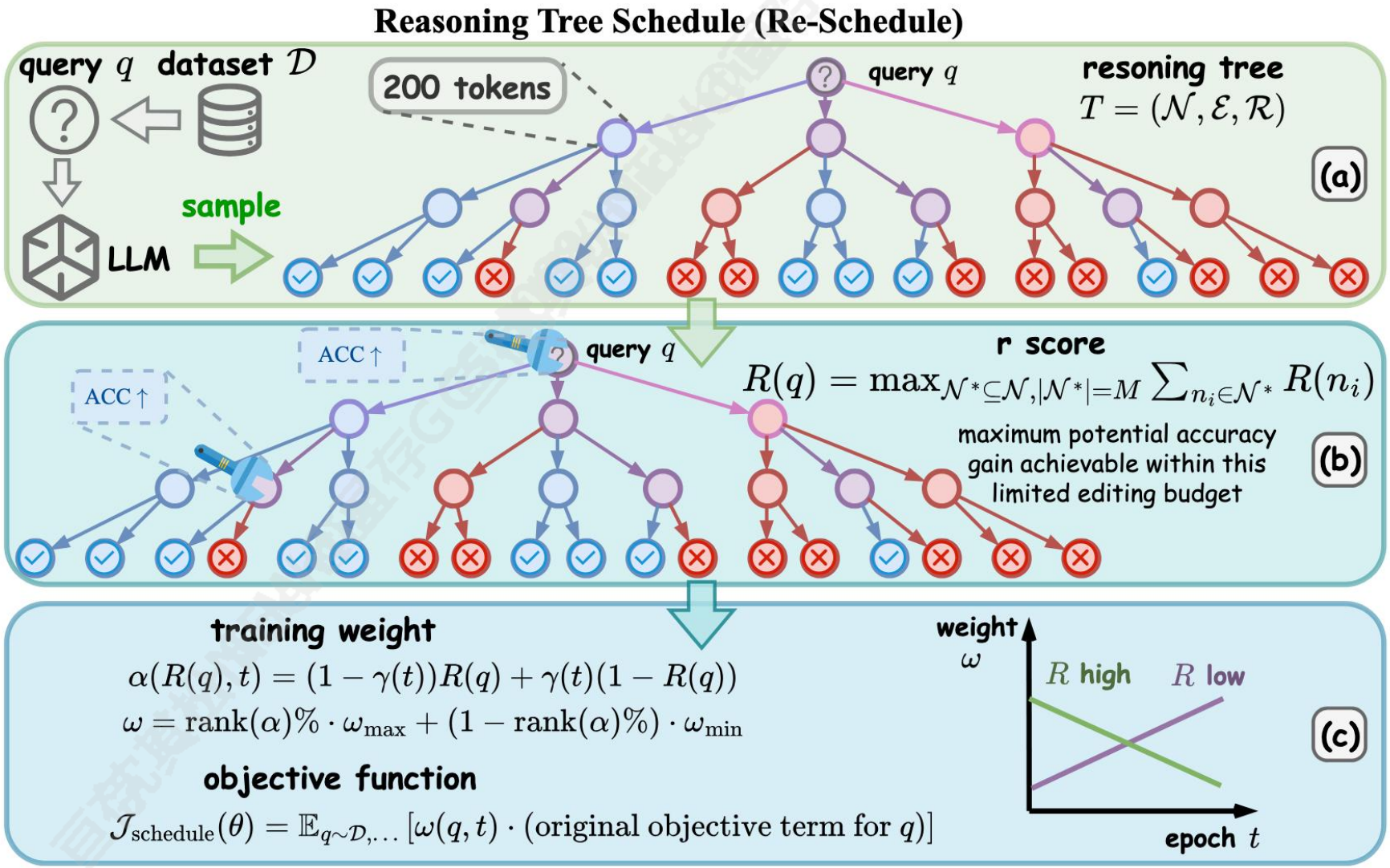


推理树结构　　最少减枝数　　关键节点数

Agentic RL中，关键节点个数、最少减枝次数 可以度量 数据训练难度

Wang H, Hao Z, Luo J, et al. Scheduling Your LLM Reinforcement Learning with Reasoning Trees. In ICLR, 2026.

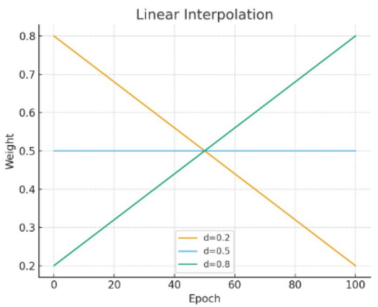
Re-Schedule 算法

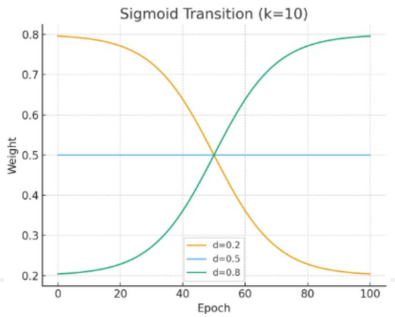


$$\text{ACC}(S) = \frac{\sum_{n_j \in S} \mathbb{I}(n_j \text{ is correct})}{|S|}$$

$$R(n_i) = \max_{n_{\text{child}} \in \mathcal{C}(n_i)} \text{ACC}\left[\mathcal{N}_{\text{leaf}} \setminus \mathcal{L}(n_i) \cup \mathcal{L}(n_{\text{child}})\right] - \text{ACC}\left[\mathcal{N}_{\text{leaf}}\right]$$

$$R(q) = \max_{\mathcal{N}^* \subseteq \mathcal{N}, |\mathcal{N}^*| = M} \sum_{n_i \in \mathcal{N}^*} R(n_i).$$

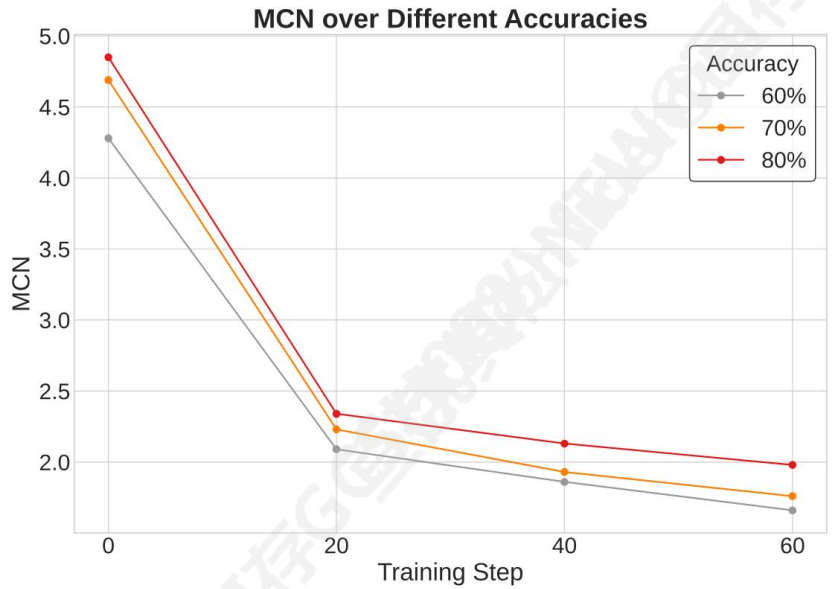




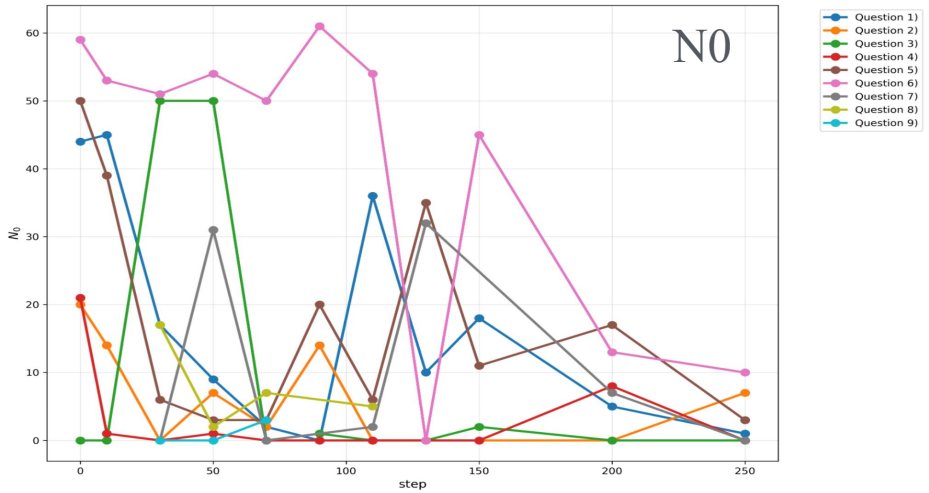
Wang H, Hao Z, Luo J, et al. Scheduling Your LLM Reinforcement Learning with Reasoning Trees. In ICLR, 2026.

实验分析

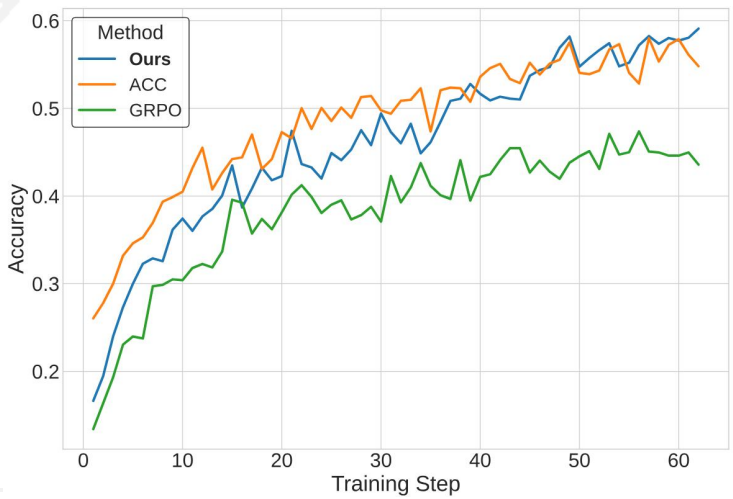
最小剪枝数

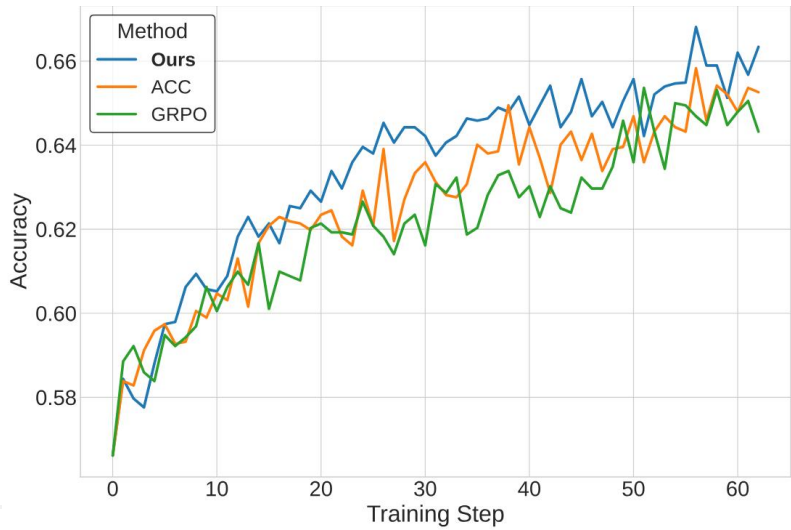


关键节点数



top 1/3 of data selected by each



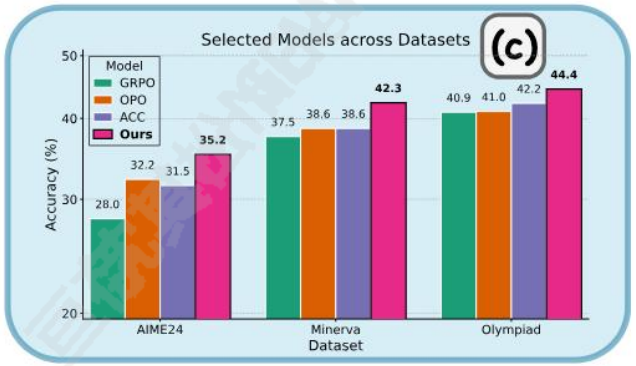


Wang H, Hao Z, Luo J, et al. Scheduling Your LLM Reinforcement Learning with Reasoning Trees. In ICLR, 2026.

训练结果

Model	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-Math-7B	13.8	5.3	44.6	39.6	9.9	13.8	21.2
Classical RLVR Methods							
GRPO	28.0	14.3	66.2	78.6	37.5	40.9	44.3
SimpleRL-Zoo	25.2	13.4	70.6	78.6	37.9	38.4	44.0
Eurus-PRIME	20.9	13.0	65.2	79.8	37.5	40.6	42.8
OPO	32.2	13.4	71.5	**82.2**	38.6	41.0	46.5
Scheduling Methods							
$ACC_{sigmoid}$	31.5	15.6	70.9	80.8	38.6	42.2	46.6
LPPO	32.8	14.9	63.3	79.2	39.0	40.6	45.0
Seed-GRPO	30.7	14.0	71.0	80.0	38.2	38.5	45.4
Our Methods							
$Re\text{-}Schedule_{linear}$	34.2	15.6	**72.4**	81.2	36.4	42.5	47.1
$Re\text{-}Schedule_{sigmoid}$	**35.2**	**16.0**	72.3	**82.2**	**42.3**	**44.4**	**48.5**

Model	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-7B	5.1	2.5	27.8	34.4	5.9	13.5	14.9
Classical RLVR Methods							
GRPO	15.6	8.8	62.5	78.2	38.6	40.4	40.7
SimpleRL-Zoo	15.7	8.8	65.7	76.0	33.8	38.7	39.8
OPO	16.6	8.4	64.6	74.6	31.6	40.3	39.4
Scheduling Methods							
$ACC_{sigmoid}$	16.7	9.8	68.6	79.0	34.2	39.4	41.3
LPPO	15.8	9.4	64.0	76.8	35.3	36.7	39.7
Seed-GRPO	13.3	6.0	63.3	76.6	32.4	36.3	38.0
Our Methods							
$Re\text{-}Schedule_{linear}$	**18.4**	12.2	68.6	80.4	41.2	42.1	43.8
$Re\text{-}Schedule_{sigmoid}$	18.2	**14.0**	**69.2**	**81.0**	**41.5**	**43.3**	**44.5**



Type	ALFWorld							WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ.
GRPO [24]	90.3	83.3	84.2	70.0	69.2	55.0	75.3	73.1	64.1
GIGPO [8]	93.5	83.3	78.9	86.7	76.2	**85.0**	83.9	83.5	75.8
HGPO [10]	92.3	91.7	77.8	93.3	85.7	73.9	85.8	88.4	77.8
Ours	92.3	**100.0**	88.9	**100.0**	76.2	82.6	90.0	89.5	80.1

Wang H, Hao Z, Luo J, et al. Scheduling Your LLM Reinforcement Learning with Reasoning Trees. In ICLR, 2026.

Agentic RL 的熵坍缩

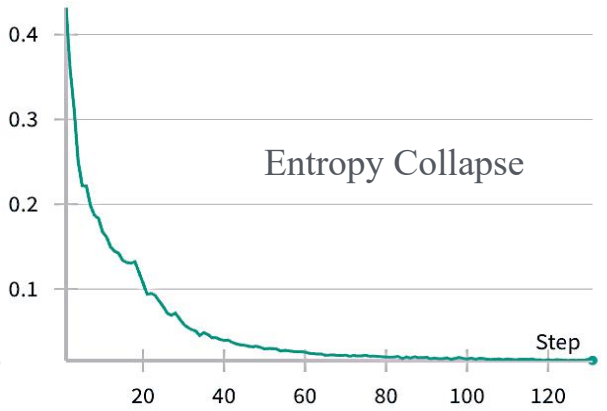
Low Entropy

High Entropy

Training Process

Token Entropy: $\mathcal{H}_{i,t} = -\mathbb{E}_{o_{i,t}\sim\pi_\theta(\cdot|q,o_{i,<t})}\left[\log \pi_\theta(o_{i,t}|q,o_{i,<t})\right]$

Policy Entropy: $\mathcal{H}(\pi_\theta,\mathcal{D}) = \mathbb{E}_{q\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\mathcal{H}_{i,t}$



Hao Z, Wang H, Liu H, et al. Rethinking entropy interventions in rlvr: An entropy change perspective. In ACL, 2026.

一系列熵干预方法

A Token-level Gradient Reweighting Perspective for Shaping Policy Entropy

- **Class 1: Clip-higher**

Method	Intervention
DAPO / DCPO (Yu et al., 2025) (Yang et al., 2025b)	$\mathbb{I}_{\text{clip}} = \begin{cases} 0, & A_{i,t} > 0 \text{ and } r_{i,t} > 1 + \varepsilon_{\text{high}}, \\ 0, & A_{i,t} < 0 \text{ and } r_{i,t} < 1 - \varepsilon_{\text{low}}, \\ 1, & \text{otherwise} \end{cases}$

- **Class 2: Positive-Reweighting**

Method	Intervention
Unlikeliness (He et al., 2025a)	$\hat{R}_{i,t} = R_{i,t}\left(1 - \beta_{\text{rank}} \frac{G - \text{rank}(o_i)}{G}\right), \ \beta_{\text{rank}} > 0$
W-REINFORCE (Zhu et al., 2025a)	$\hat{A}_{i,t} = \begin{cases} \lambda, & A_{i,t} > 0 \\ 1, & A_{i,t} < 0 \end{cases}, \quad \lambda < 1$

- **Class 3: Entropy-Aware Advantage**

Method	Intervention
Entropy Advantage (Cheng et al., 2025)	$\hat{A}_{i,t} = A_{i,t} + \min\left(\alpha \cdot \mathcal{H}_{i,t}^{\text{detach}}, \frac{\lvert A_{i,t}\rvert}{\kappa}\right), \quad \alpha > 0, \kappa > 1$
GTPO (Tan & Pan, 2025)	$\hat{R}_{i,t} = R_{i,t} + \alpha \frac{\mathcal{H}_{i,t}}{\frac{1}{d_t}\sum_{k=1}^{d_t} \mathcal{H}_{k,t}}, \text{ for } R_{i,t} > 0$

Why and how they work ?

Policy Gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\ \{o_i\}\sim\pi_{\text{old}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} w_{i,t}(q)\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t})\right]$$

Token Gradient Reweight:

$$w_{i,t}(q) = \mathbb{I}_{\text{clip}} r_{i,t} A_{i,t} + \beta \mathcal{R}(\pi_\theta)$$

$$\mathbb{I}_{\text{clip}} = \begin{cases} 0, & A_{i,t} > 0 \text{ and } r_{i,t} > 1 + \varepsilon, \\ 0, & A_{i,t} < 0 \text{ and } r_{i,t} < 1 - \varepsilon, \\ 1, & \text{otherwise}, \end{cases}$$

Hao Z, Wang H, Liu H, et al. Rethinking entropy interventions in rlvr: An entropy change perspective. In ACL, 2026.

token熵变的视角

Entropy Dynamics → Token-level Entropy Change

The Estimated Entropy Change:

$$\Omega(s) \triangleq -\frac{\eta}{L}\mathbb{E}_{\pi_\theta(\cdot|s)}\left[\frac{\mathbb{I}_{\text{clip}}\,A}{\pi_{old}}\pi_\theta(1-\pi_\theta)\left(\log\pi_\theta+\mathcal{H}(s)\right)\right]$$

$$\delta(\pi_\theta,\mathcal{H}) \triangleq -\pi_\theta(1-\pi_\theta)\left[\log(\pi_\theta)+\mathcal{H}(s)\right]$$

Remark 1: Accurate Estimator

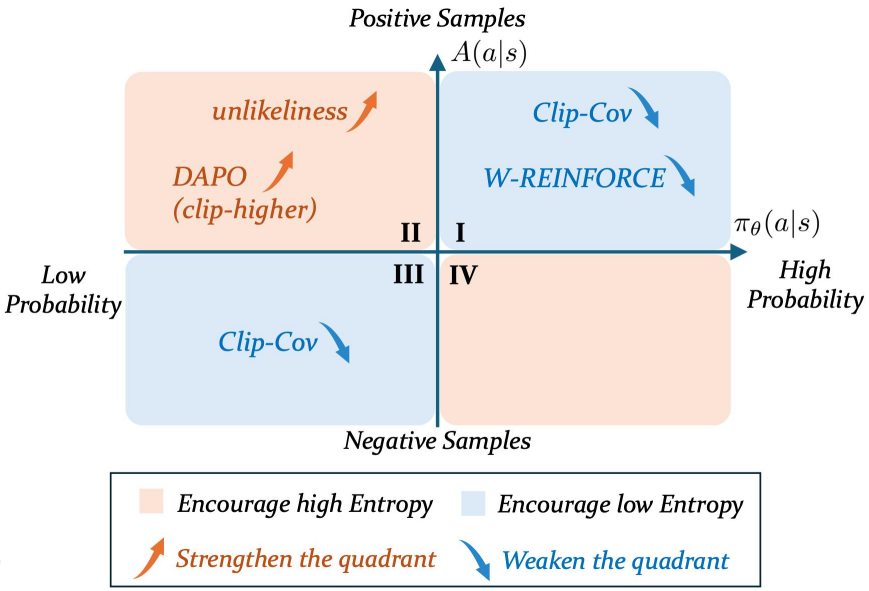
Model	Method	MSE↓	PCC↑	SRCC↑
Math-1.5B	Cov	5.37	−6e−5	+0.04
	Ours	**5e−4**	**+0.42**	**+0.65**
Qwen-7B	Cov	0.53	+0.05	+0.08
	Ours	**8e−4**	**+0.39**	**+0.72**
Math-7B	Cov	0.29	+0.03	+0.06
	Ours	**4e−4**	**+0.42**	**+0.61**

Table 1: Accuracy of entropy change estimation: the baseline *Cov* vs. our estimator.

Remark 2: Four Governing Factors

- **Clipping** 裁剪策略
- **Advantage** 优势信号
- **π** Token 概率
- **H** 条件熵

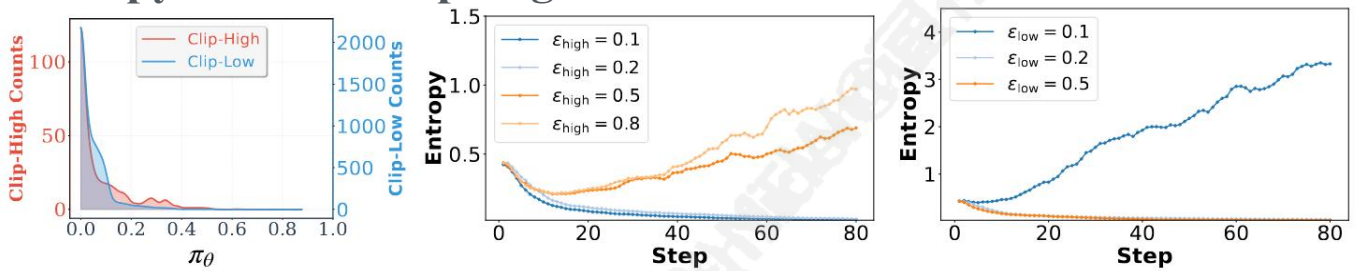
Remark 3: Entropy Change Direction



Hao Z, Wang H, Liu H, et al. Rethinking entropy interventions in rlvr: An entropy change perspective. In ACL, 2026.

对现有方法的一些解释

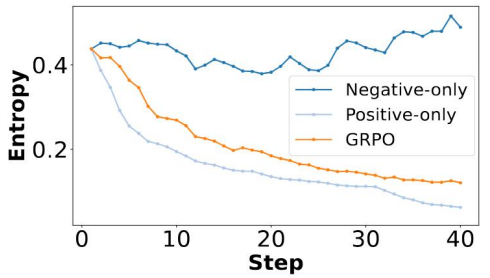
Entropy Effect of Clip-Higher



(a) Clipping statistics (b) Entropy effect of ε_{high} (c) Entropy effect of ε_{low}

Clip-high contributes to entropy increase, while ***Clip-low*** contributes to entropy decrease.

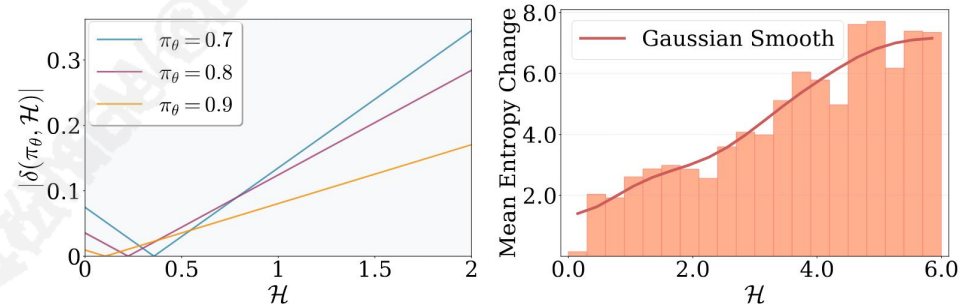
Entropy Effect of Re-weighting Positives



(d) Positive/Negative-only

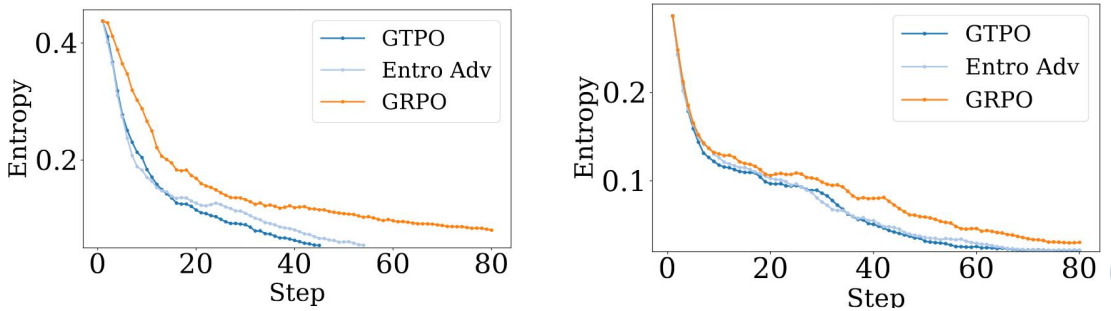
Training on ***positive samples*** leads to entropy decrease , while training on ***negative samples*** leads to entropy increase.

Entropy Effect of Entropy-aware Advantage



(a) Theoretical correlation (b) Empirical behavior

Figure 4: Entropy change as a function of entropy.



(a) On dataset DAPO-17k (b) On dataset Math

Figure 5: Entropy dynamics with entropy-aware advantage on Qwen2.5-Math-7B backbone.

High-entropy tokens tend to induce larger entropy changes

熵稳定方法 STEER

The design of STEER

Adaptive token reweighting: $\Omega(s) \rightarrow \lambda(s)\Omega(s)$

Reweighting by a simple exponential decay function:

$$\lambda(s) = \exp\left(-\alpha \frac{|\Omega(s)|}{\max_{s \in \mathcal{B}} |\Omega(s)|} \right)$$

$\alpha > 0 \longrightarrow$ the hyperparameter controlling the decay rate

Hyperparameter: $\lambda_{\min} = \exp(-\alpha) \longrightarrow \lambda(s) \in [\lambda_{\min}, 1]$

The advantage of STEER

- monotonically decreasing with the normalized entropy change
- all weights remain strictly positive
- the weights of most tokens are 1
- only one hyperparameter

Method	$\mathbb{I}_{\text{clip}}$	π_θ	A	$\mathcal{H}(s)$
DAPO	✓	✗	✗	✗
Unlikeliness	✗	✓	✓	✗
W-REINFORCE	✗	✗	✓	✗
Entro. Adv.	✗	✗	✓	✓
KL Reg.	✗	✓	✗	✗
Entropy Reg.	✗	✗	✗	✓
Edge-GRPO	✗	✗	✗	✓
Forking Tokens	✗	✗	✗	✓
Clip-Cov	✗	✓	✓	✗
STEER	✓	✓	✓	✓

Table 2: Factors that governing entropy change considered by existing entropy-intervention methods and ours.

Existing strategies involve qualitative adjustments to only one or two factors, while STEER includes all factors.

Hao Z, Wang H, Liu H, et al. Rethinking entropy interventions in rlvr: An entropy change perspective. In ACL, 2026.

训练结果

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-Math-7B	13.8	5.3	44.6	39.6	9.9	13.8	21.2
Classical RLVR Methods							
GRPO	28.0	14.3	66.2	78.6	37.3	40.9	44.2
SimpleRL-Zoo	30.8	14.2	65.4	79.2	37.1	40.8	44.6
Eurus-PRIME	20.9	13.0	65.2	79.8	37.4	40.6	42.8
OPO	32.2	13.4	71.5	82.2	38.2	41.0	46.4
Entropy Intervention Methods							
GRPO w/ clip-high	31.7	12.8	66.8	79.0	38.6	39.3	44.7
GRPO w/ Entro. Loss	29.1	14.0	67.6	80.0	38.2	37.9	44.5
GRPO w/ Fork Tokens	31.9	14.3	65.5	79.2	37.1	40.9	44.8
W-REINFORCE	29.2	12.0	64.9	79.2	37.8	40.9	44.0
Entro. Adv.	27.5	13.5	70.2	79.6	36.8	42.8	45.1
Clip–Cov	32.5	12.9	68.4	78.0	40.8	41.3	45.7
KL–Cov	32.8	14.1	64.2	78.8	37.1	39.4	44.4
STEER (Ours)	**36.2**	**16.1**	**72.1**	**82.2**	**41.7**	**43.0**	**48.6**

Table 3: Benchmark results of different methods on Qwen2.5-Math-7B. We report avg@32 for AIME24, AIME25, and AMC23 and avg@1 for other datasets. All results are presented as percentages.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
GRPO	31.6	12.8	66.7	79.0	39.3	40.1	44.9
Entro. Adv.	34.8	13.4	64.3	77.6	37.6	39.9	44.6
Entro. Loss	32.7	14.7	71.3	79.0	36.8	41.4	46.0
Clip-Cov	30.4	14.0	72.3	79.6	37.1	41.7	45.8
STEER (Ours)	**36.1**	**16.0**	**76.3**	**80.5**	**39.5**	**42.3**	**48.5**

Table 6: Performance on test datasets in extreme scenarios.

Dataset	Method	3B	7B	14B
Internal	GRPO	39.7	40.1	42.6
	STEER	**41.2**	**41.9**	**45.1**
Zeta	GRPO	17.4	22.0	22.3
	STEER	**19.3**	**24.0**	**24.1**
LCB-v5	GRPO	24.4	28.5	29.3
	STEER	**24.9**	**29.2**	**31.8**

Table 5: Performance on real-world coding tasks.

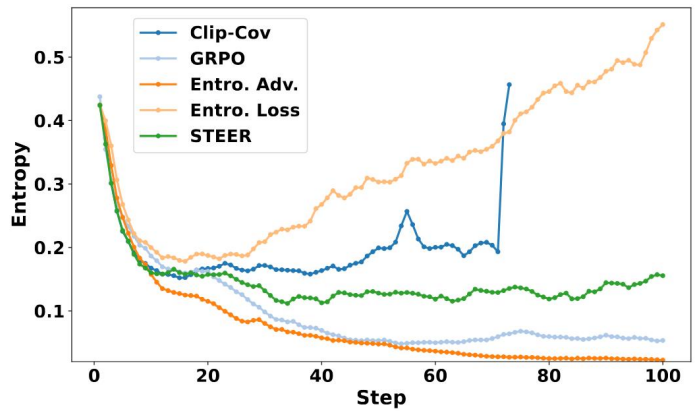


Figure 8: Entropy dynamics in extreme scenarios.

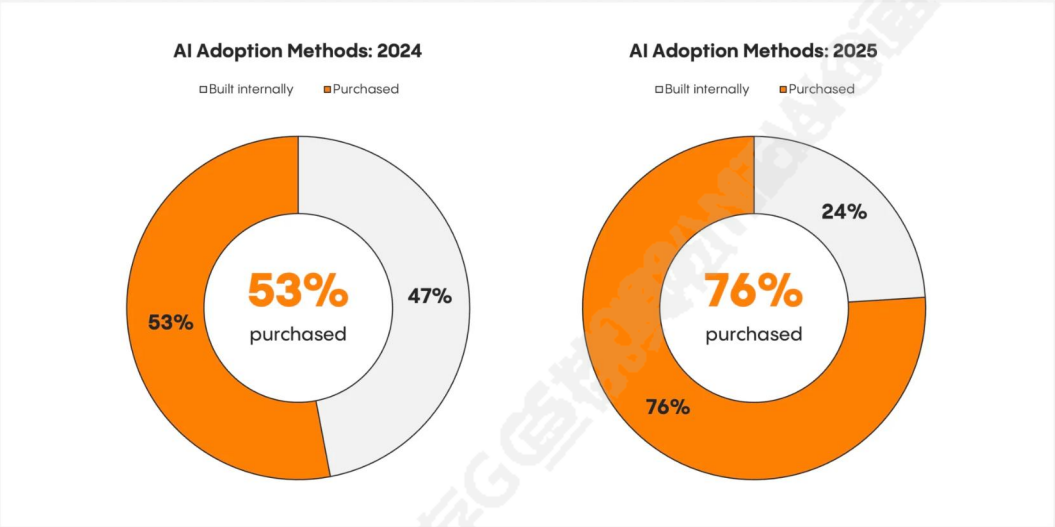
Hao Z, Wang H, Liu H, et al. Rethinking entropy interventions in rlvr: An entropy change perspective. In ACL, 2026.

Code Agent 如何在训练-测试-部署的闭环中持续提升能力?

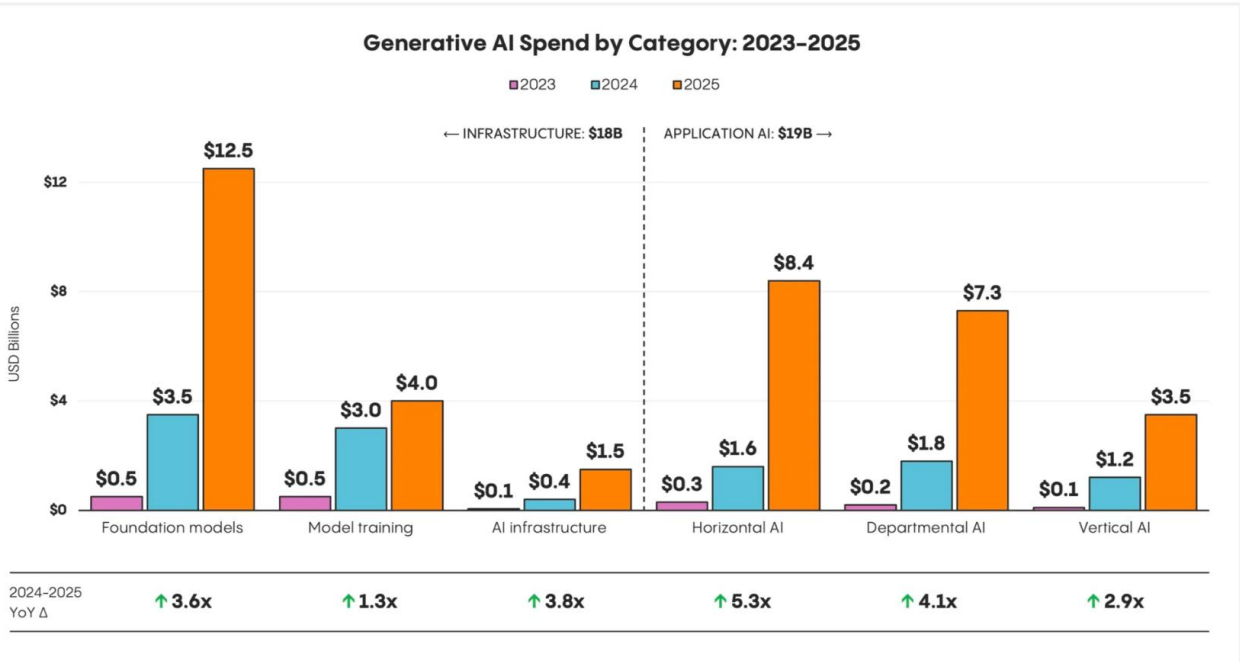
目录

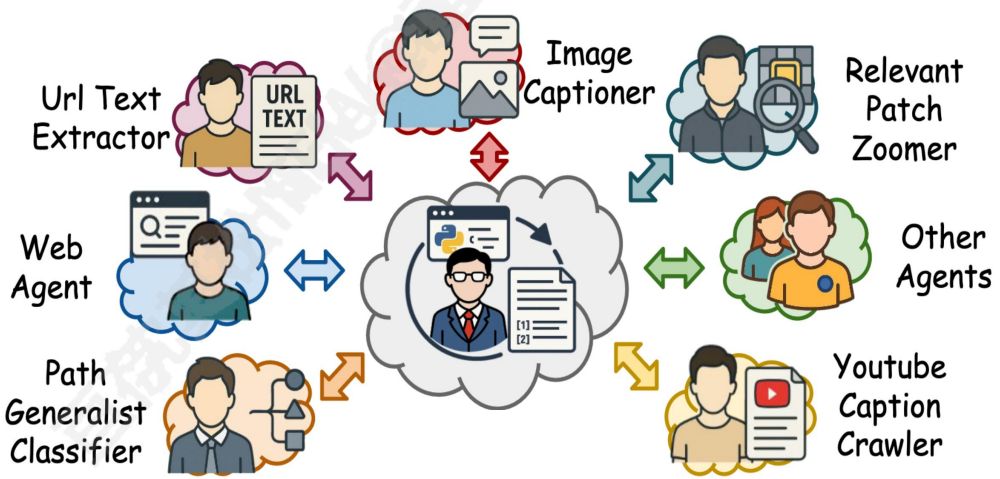
Agents Development

Building vs. Buying Enterprise AI Solutions



Where Does the GenAI Budget Go?





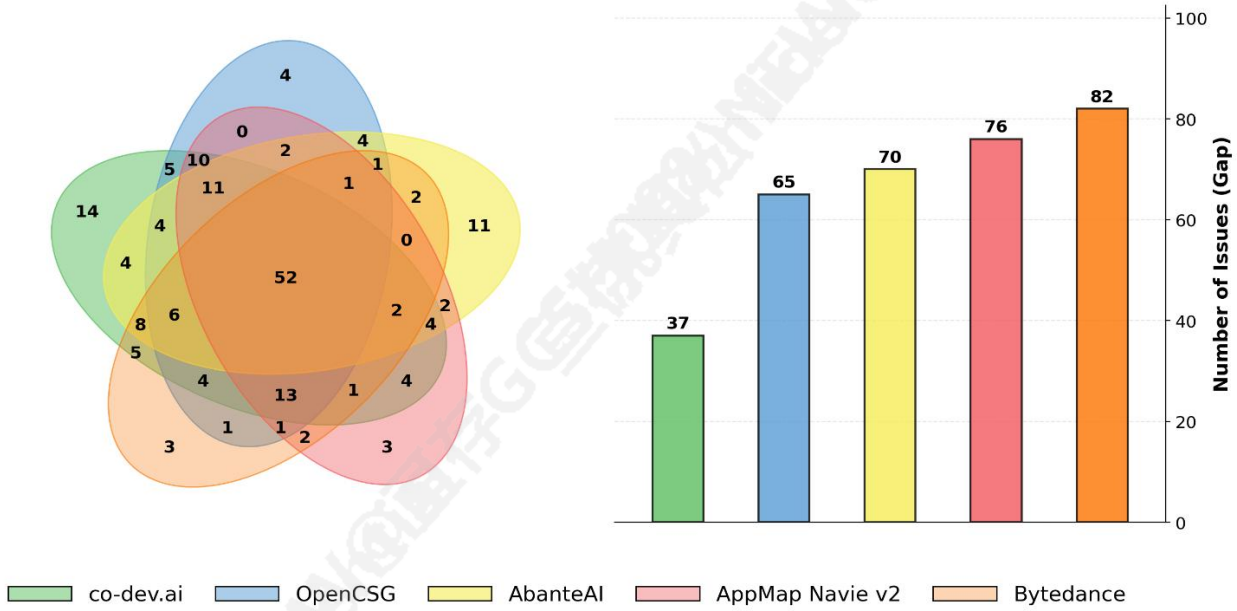
Large-scale Manual Engineering

自动地从零开始，
为任意新领域创建专业 Agent?

Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

如何高效地利用执行经验？

Human-designed Scaffold



SWE-bench Lite (300 cases)

Pass rate ranges from 134 to 187

Current agent scaffolds gate what a base model can do.

Experience Matters

Experience can suggest adding rules, creating tools, modifying workflows…

Data Science

(no train/test split)
clf.fit(X, y);
clf.score(X, y)

→ add rules:
construct train/test split

Software Engineering

(empty patch)
git diff path/to/file.py
git commit
[complete]

→ edit workflow:
6. git diff –cached
7. [complete]

Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

面向Code Agent的演化方法：ReCreate

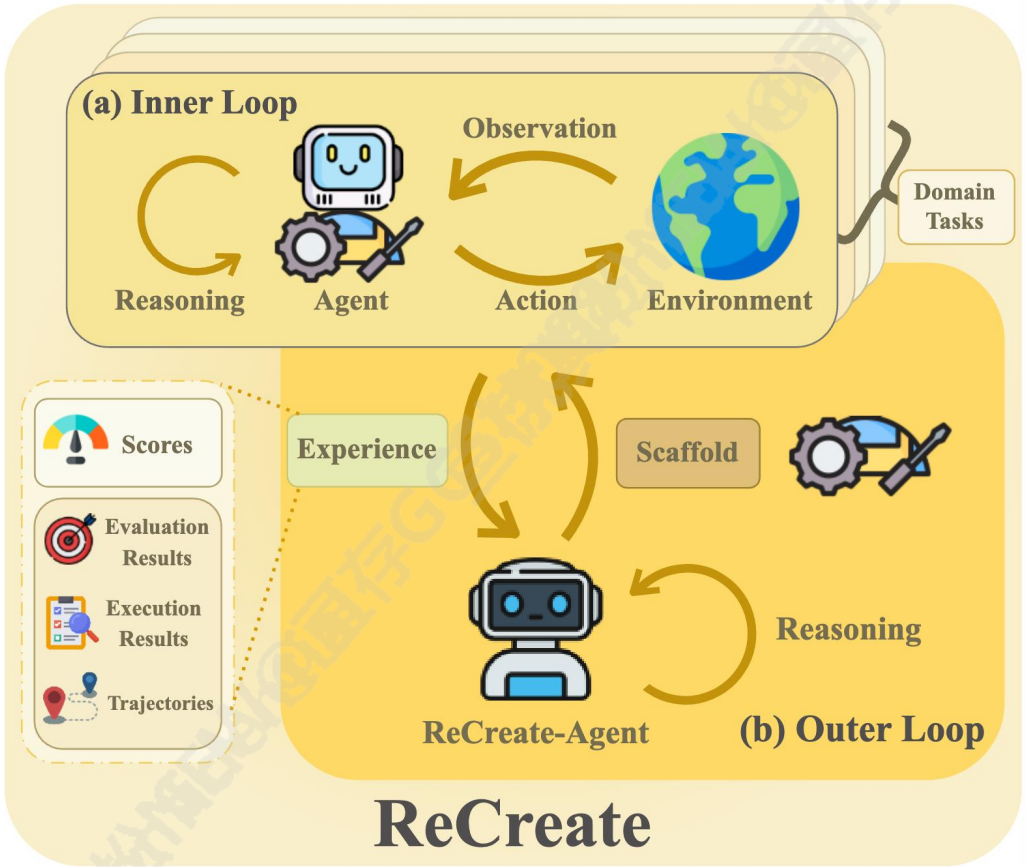


Figure 2: The overview of ReCreate.

Optimization:

$$\max_{\mathcal{A}} \ \mathbb{E}_{t_i \sim \mathcal{D}} \mathrm{Ver}\left[\mathrm{Exec}\left(\tau_i^*(\mathcal{A}, t_i), t_i\right)\right]$$
$$\text{s.t. } \tau_i^*(\mathcal{A}, t_i) \in \arg \max_{\tau_i \sim P_\phi(\cdot|\mathcal{A}, t_i)} \mathrm{Ver}\left(\mathrm{Exec}(\tau_i, t_i)\right).$$

Automated-Agent Design:

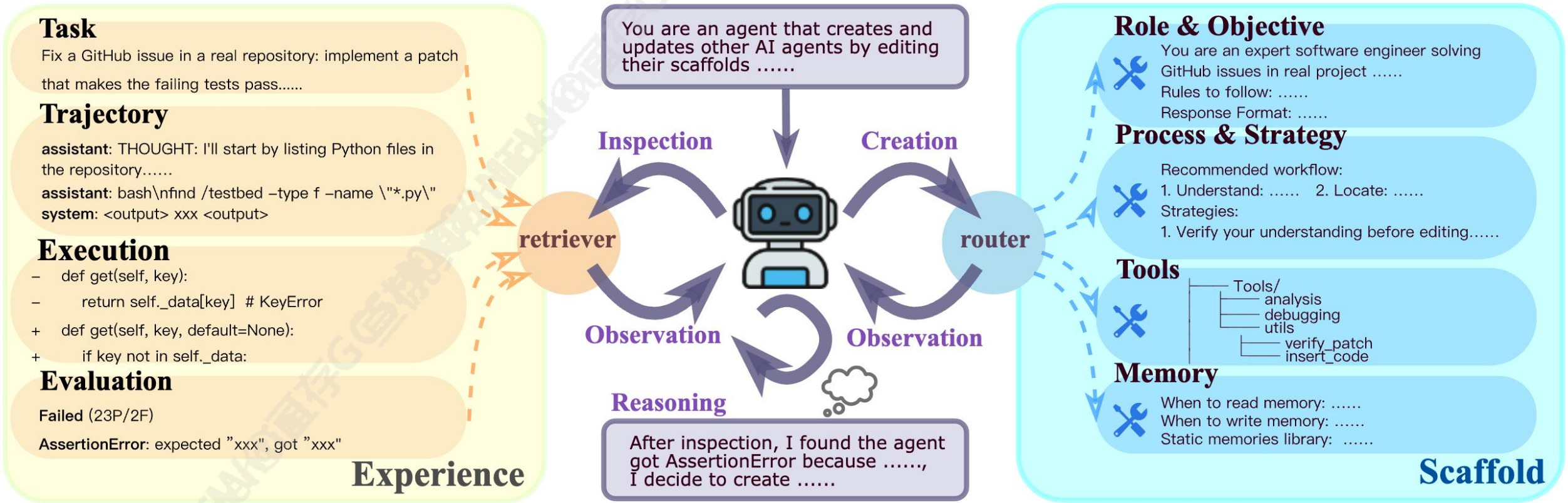
$$\mathcal{A}_{k+1} = \text{Meta-Agent}\left(\mathcal{A}_k, r\right)$$

ReCreate:

$$\mathcal{A}_{k+1} = \text{ReCreate-Agent}\left(\mathcal{A}_k, e\right)$$
$$\text{where } e \triangleq \left(\tau,\ \mathrm{Exec}, \mathrm{Ver}\right).$$

Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

面向Code Agent的演化方法：ReCreate



The Agent-as-optimizer Design

- 完整的交互经验过长，难以处理和分析
- 从完整的交互经验中归因并转化为有效的修改存在挑战
- 单个经验可能难以泛化，导致Agent越学越偏

Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

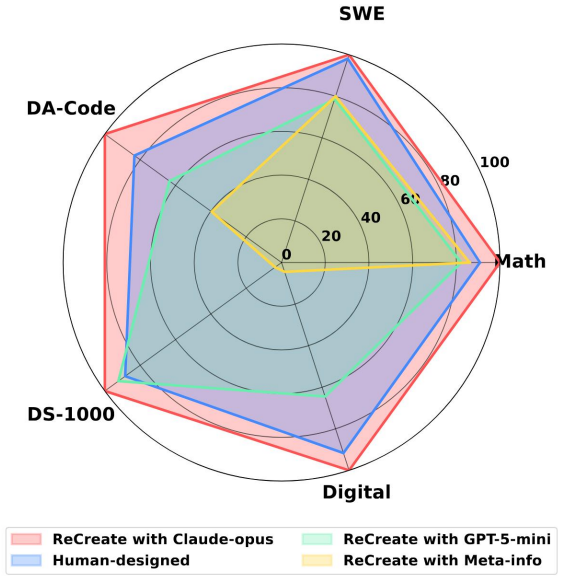
多场景下的演化结果

Domain	Dataset	Human-designed			Self-Evolve			Agent Generation		Ours
		CoT	SBA	Refine	LIVE	OPRO	ExpeL	ADAS	Square	ReCreate
SWE	*Django*	58.29	58.77	56.87	58.77	52.13	59.24	55.92	57.82	**60.19**
	Sympy	61.82	61.82	58.18	58.18	54.55	56.36	58.18	54.55	**63.64**
DS	*DW*	42.81	41.66	42.43	38.55	18.55	47.50	37.98	45.98	**51.94**
	ML	34.32	36.25	35.85	41.68	39.08	40.27	21.53	22.90	**42.88**
	SA	19.33	20.15	20.39	21.83	17.16	22.44	15.66	11.99	**24.50**
	Numpy	62.00	64.50	64.00	68.00	71.00	68.50	61.00	67.00	**77.00**
	Pandas	62.73	61.25	63.10	64.21	63.47	64.94	60.52	63.10	**68.63**
	Matplotlib	78.52	80.74	81.48	82.96	82.96	78.52	76.30	82.96	**85.19**
Math	*Algebra*	81.45	85.48	84.68	85.48	87.10	83.87	80.65	83.87	**92.74**
	NT	91.94	90.32	90.32	93.55	**100.00**	90.32	83.87	93.55	**100.00**
	C&P	94.74	94.74	94.74	92.11	94.74	92.11	84.21	94.74	**100.00**
Digital	*Normal*	48.81	48.81	47.02	47.62	51.79	49.40	50.00	48.81	**52.98**
	Challenge	34.05	36.21	35.01	34.53	34.29	34.53	36.93	34.77	**40.29**
Average		59.29	60.05	59.54	60.57	58.99	60.62	55.60	58.62	**66.15**

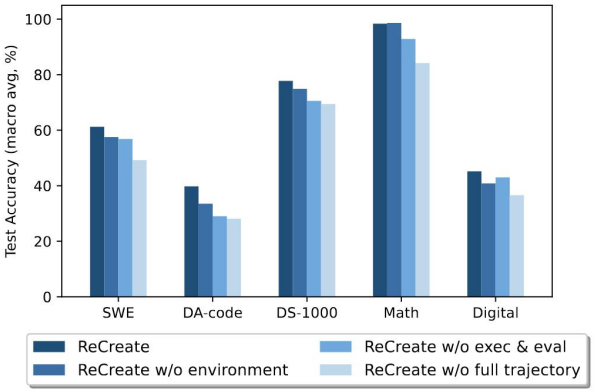
Main Results

Method	SWE	DA-Code	DS
ReCreate	60.19	39.77	77.74
w/o creation router	58.00	37.13	75.39
w/o DOMUPD	57.09	37.83	75.96

Action-level Ablation



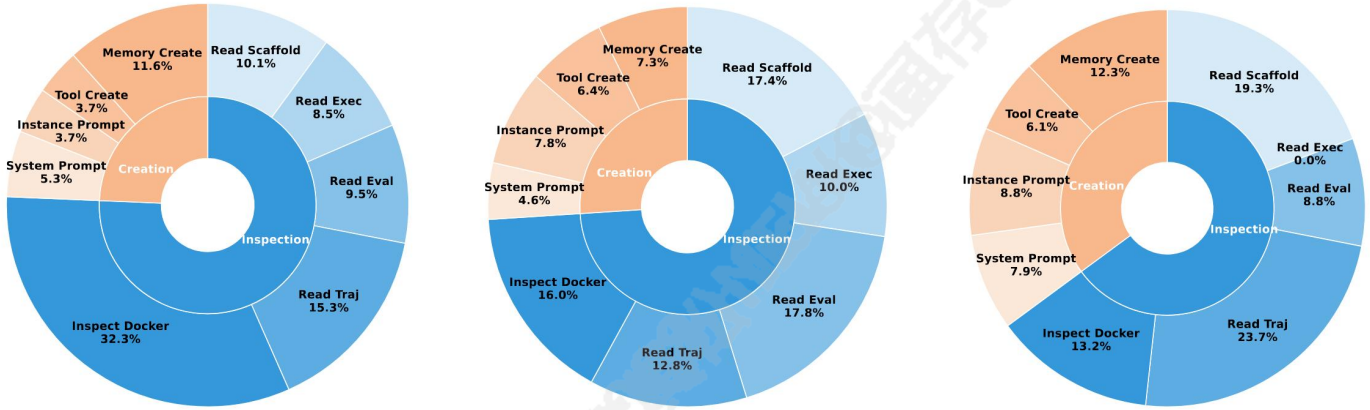
Reasoning-level Ablation



Observation-level ablation

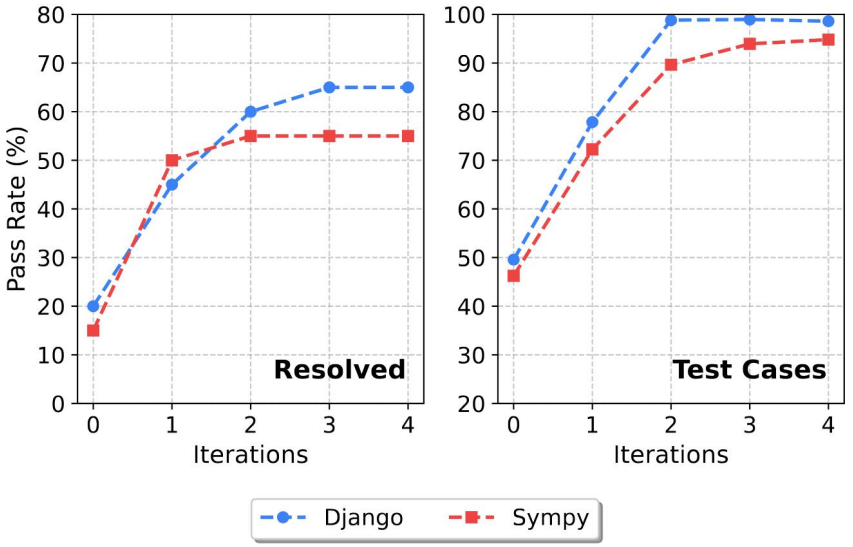
Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

演化过程分析

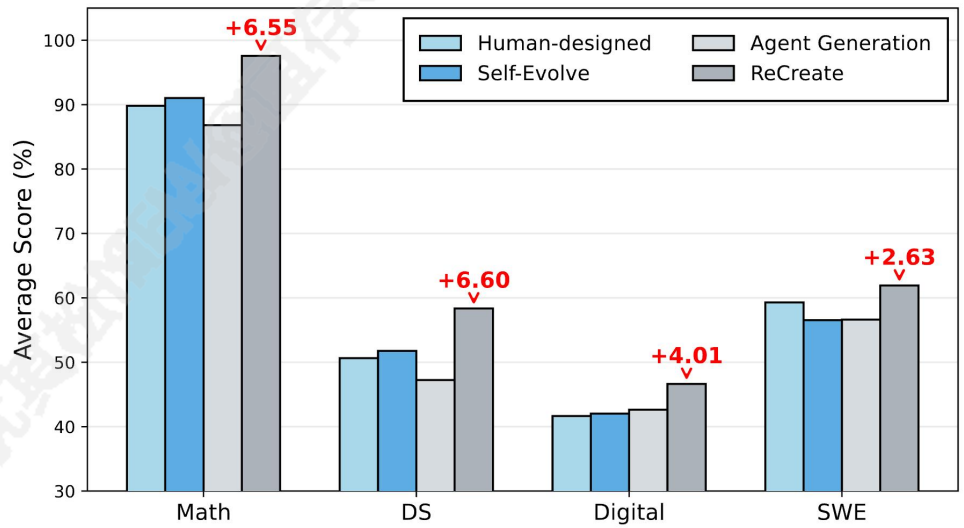


(a) Django (b) Machine Learning (c) AppWorld

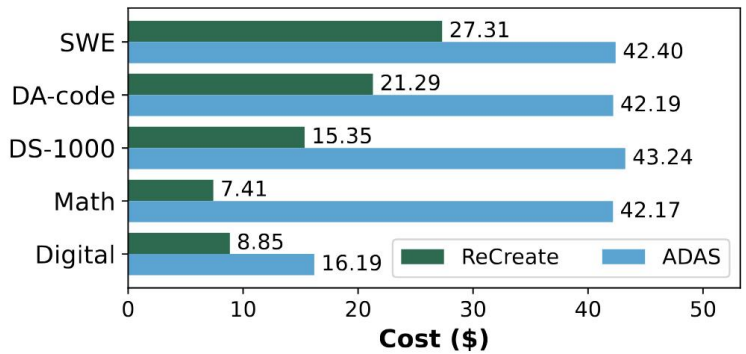
Statistical Analyses



Evolving Process



Overall Performance

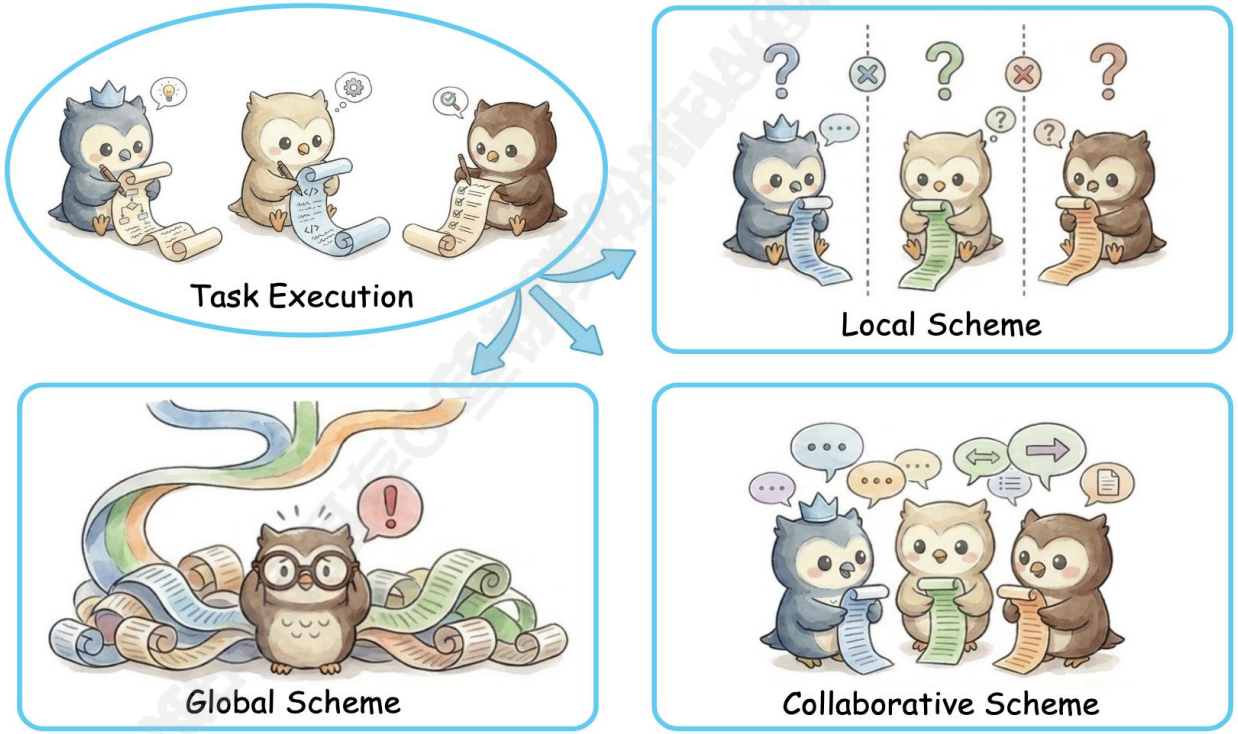


Cost Comparison

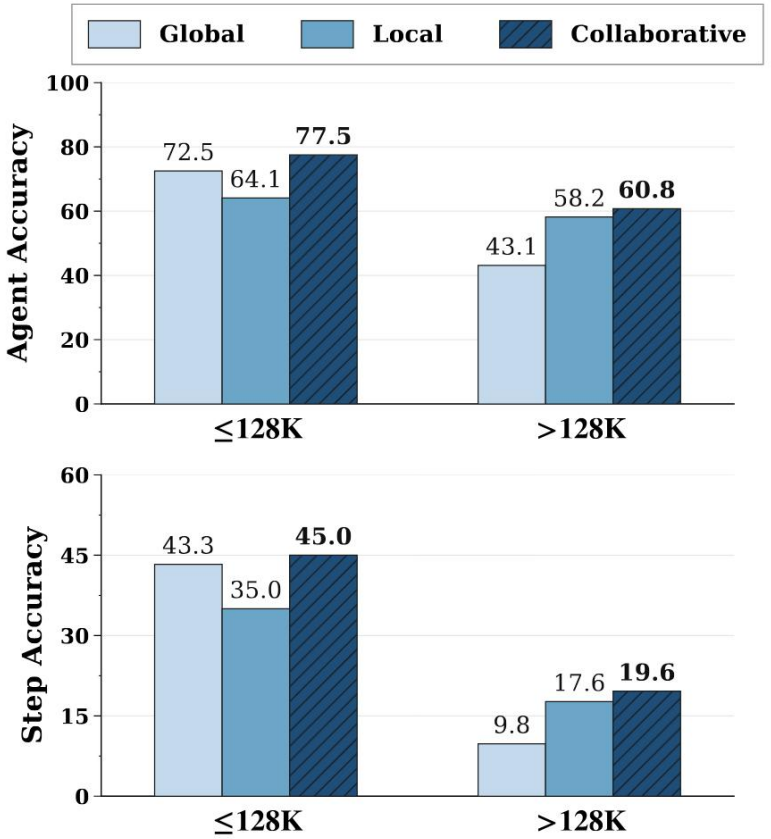
Hao Z, Wang H, Luo J, et al. ReCreate: Reasoning and Creating Domain Agents Driven by Experience. In ACL, 2026.

Agent Team的演化：Evolve as a Team

通过协作演化解决 整体—局部 演化的困境



(a) Schematic illustration.

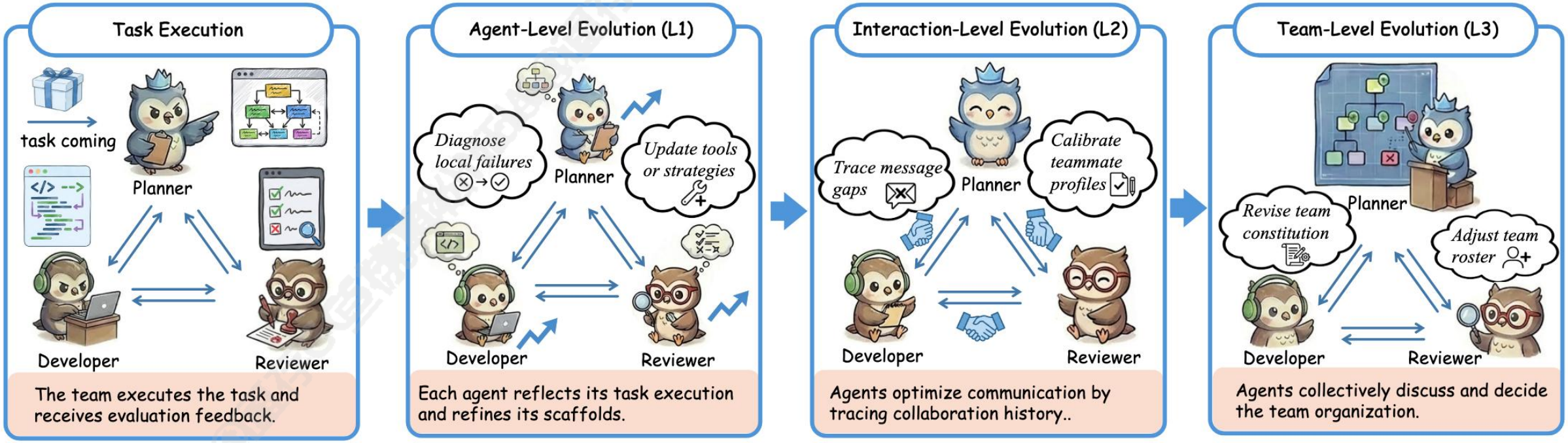


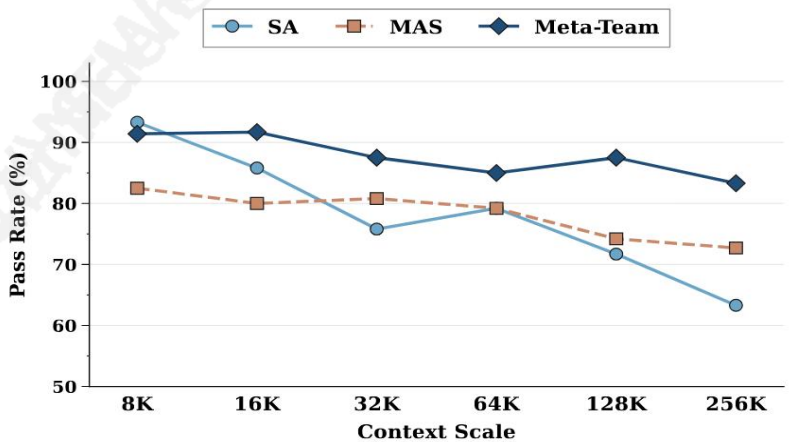
(b) Attribution comparison.

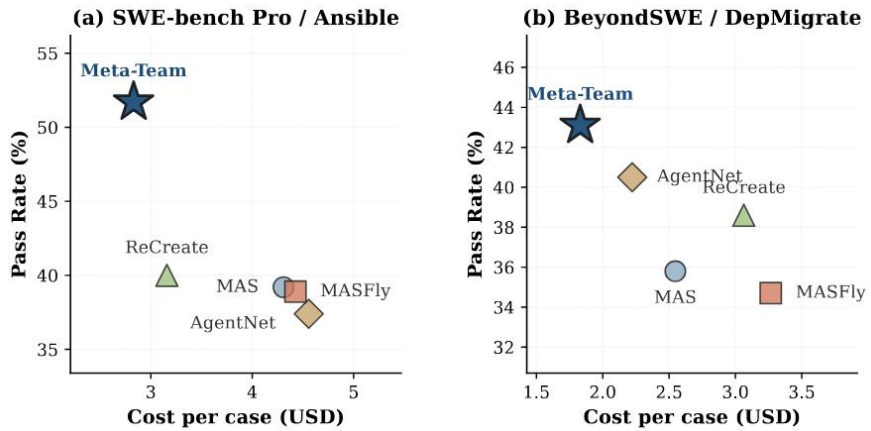
- 提供最终结果之外的反馈信号
- 去中心化的演化决策机制

Hao Z, Wang T, Dong H, et al. Evolve as a Team: Collaborative Self-Evolution for LLM-based Multi-Agent Systems.

Meta-Team 演化框架







Hao Z, Wang T, Dong H, et al. Evolve as a Team: Collaborative Self-Evolution for LLM-based Multi-Agent Systems.

演化测试

Method	SWE-Pro		BeyondSWE		LOCA	GAIA	LoCoBench		ResRub.	Avg.
	Ansible	Qute.	DepMig.	CrossR.	Val	Val	Feat.	Refact.	All	
SA	44.7	62.7	43.5	41.7	78.3	70.0	64.5	62.0	43.9	56.8
MAS	40.8	61.0	37.6	40.4	82.1	74.7	57.6	60.9	49.5	56.1
AggAgent	49.6	57.7	42.8	40.6	80.4	70.3	60.2	54.7	47.2	55.9
OWL	39.5	60.1	36.7	36.9	72.9	64.0	53.8	56.2	47.9	52.0
AOrchestra	44.7	54.0	44.1	42.0	82.9	73.3	62.0	65.5	51.5	57.8
AgentSquare	39.0	56.3	38.4	37.4	68.3	65.0	56.2	55.9	41.9	50.9
ReCreate	42.5	59.6	41.6	39.1	74.6	69.0	58.9	62.0	45.9	54.8
AgentNet	40.8	58.7	44.3	39.6	69.6	67.3	57.0	56.9	40.3	52.7
MASFly	43.4	62.9	39.7	39.4	75.8	71.0	60.1	64.5	50.8	56.4
Meta-Team	**53.9**	**66.2**	**45.6**	**43.3**	**87.9**	**77.3**	**67.1**	**67.0**	**55.8**	**62.7**

Abbreviations. Ans. = ansible, Qute. = qutebrowser, DepMig. = DepMigrate, CrossR. = CrossRepo, Feat. = Feature Implementation, Refact. = Cross-File Refactoring, ResRub. = ResearchRubrics. **Avg.** is the unweighted arithmetic mean across the nine columns.

Results across diverse agent benchmarks

Experience schemes	Ansible ↑	ResRub. ↑
No evolution (MAS)	40.8	49.5
Centralized experience	49.8	52.9
Partitioned experience	44.5	51.0
Collaborative experience	**53.9**	**55.8**

Ablation on collaborative self-evolution.

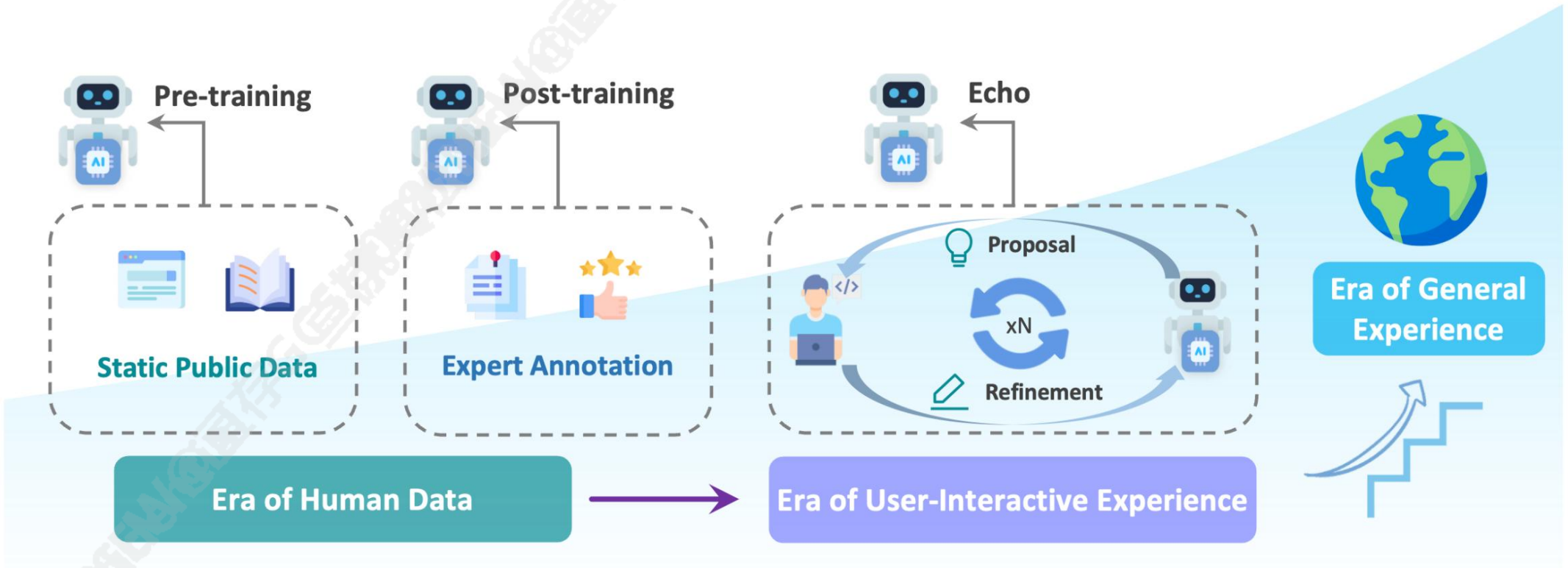
Variant	Ansible ↑	ResRub. ↑
Full Meta-Team	**53.9**	**55.8**
w/o L1 (agent)	48.5	47.9
w/o L2 (interaction)	50.7	54.4
w/o L3 (team)	51.9	52.1

Ablation on multi-scale evolution

Code Agent 如何在训练-测试-部署的闭环中持续提升能力?

目录

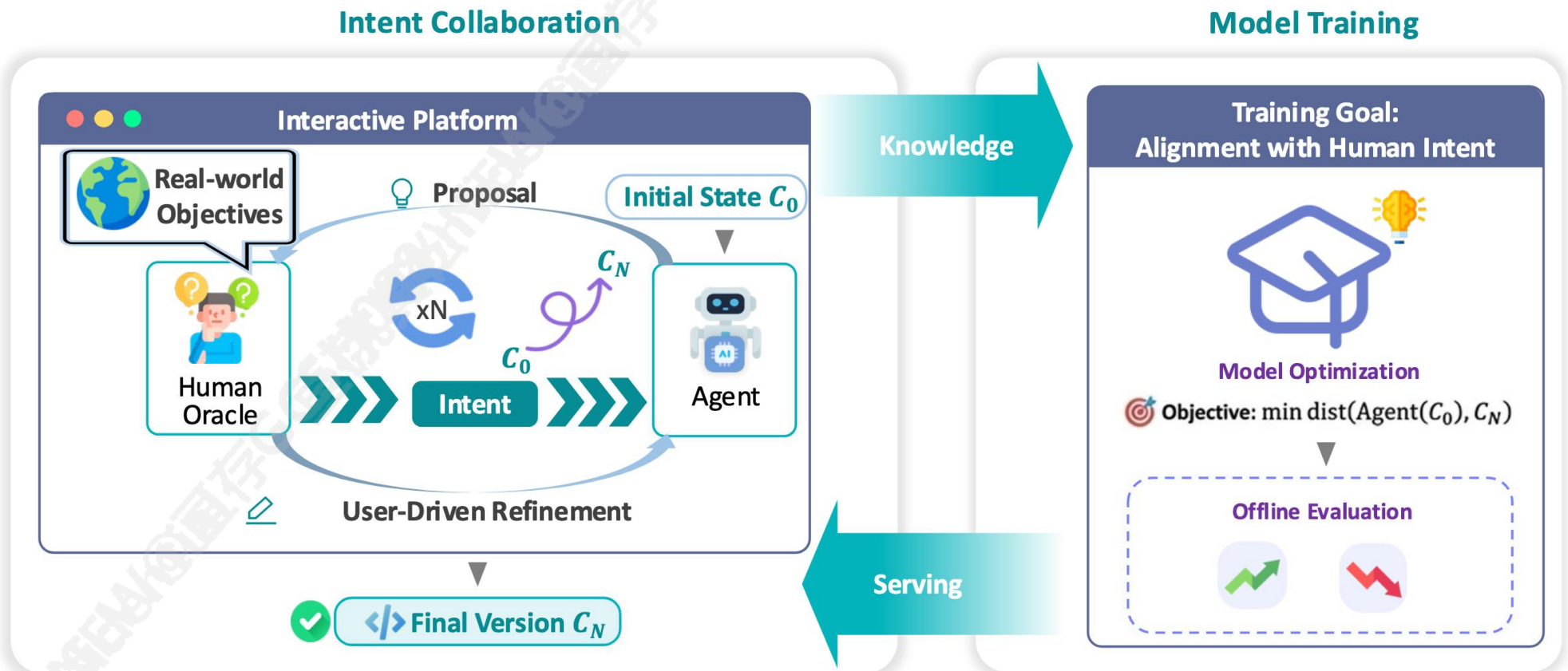
范式转变：从静态存档到交互经验



Raw experience → Learnable knowledge

Dong H, Liang X, Yu J, et al. Echo: Learning from Experience Data via User-Driven Refinement.

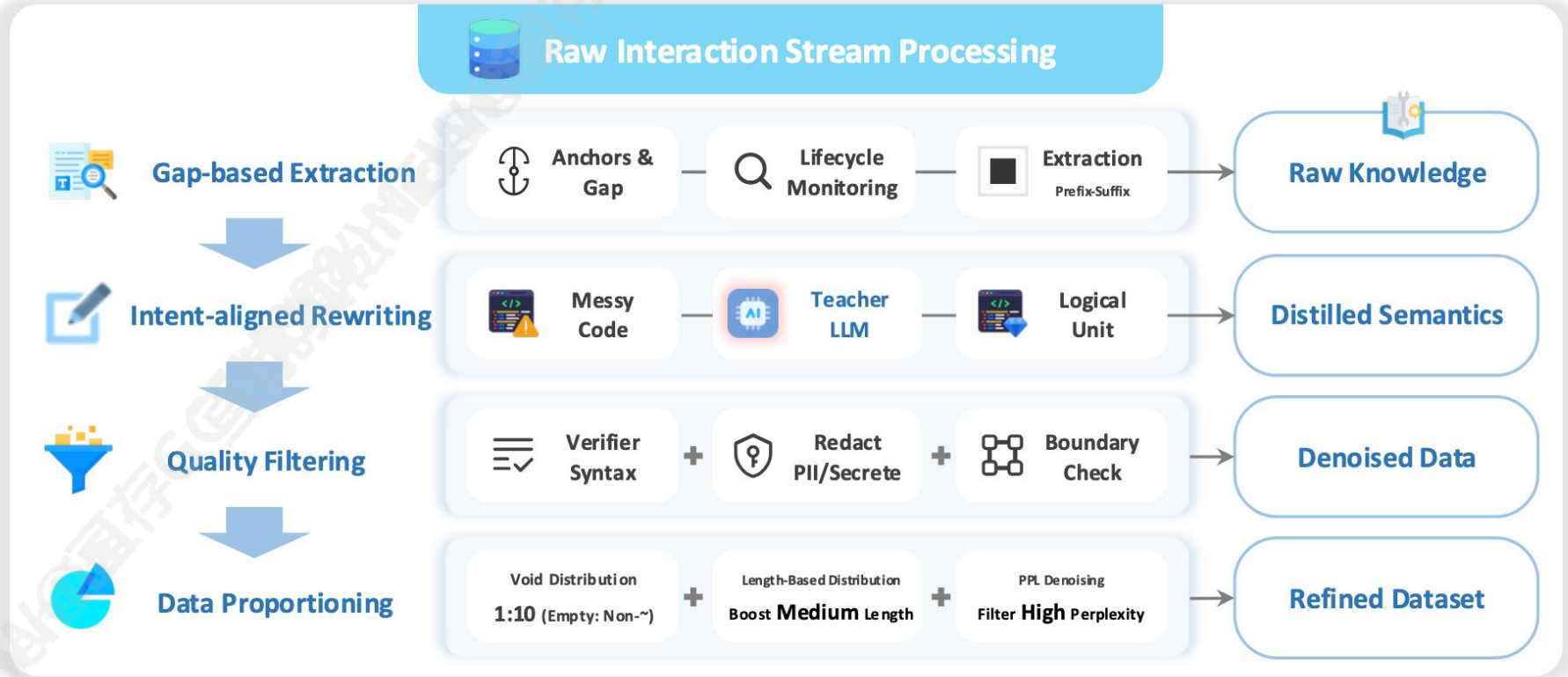
Echo 框架



$$\text{State } C_0 \xrightarrow{\text{Agent}} \underbrace{C_1 \xrightarrow{\text{User Interventions}} C_N}_{\textbf{Refinement Sequence}}$$

Dong H, Liang X, Yu J, et al. Echo: Learning from Experience Data via User-Driven Refinement.

Echo 框架 – 补全模型的尝试

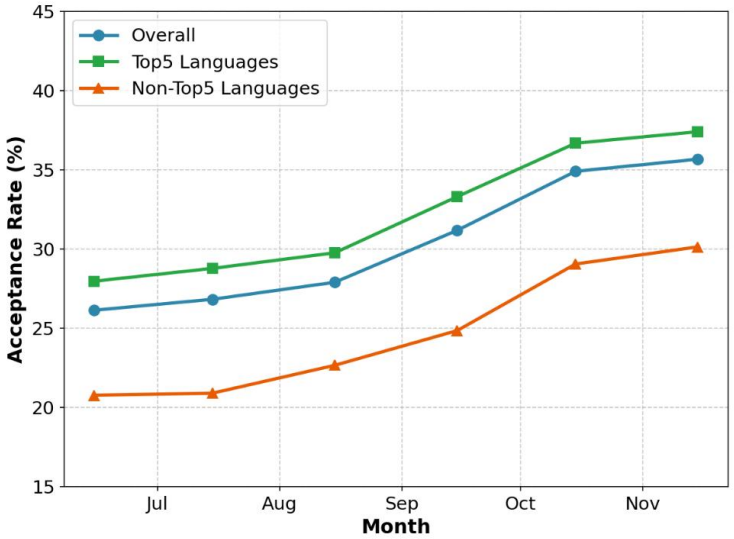


$$\mathcal{L}_{echo}(\theta) = \mathbb{E}_{(C_0, C_N)\sim\mathcal{D}_{echo}}[\mathrm{dist}(\mathrm{Agent}_\theta(C_0), C_N)],$$

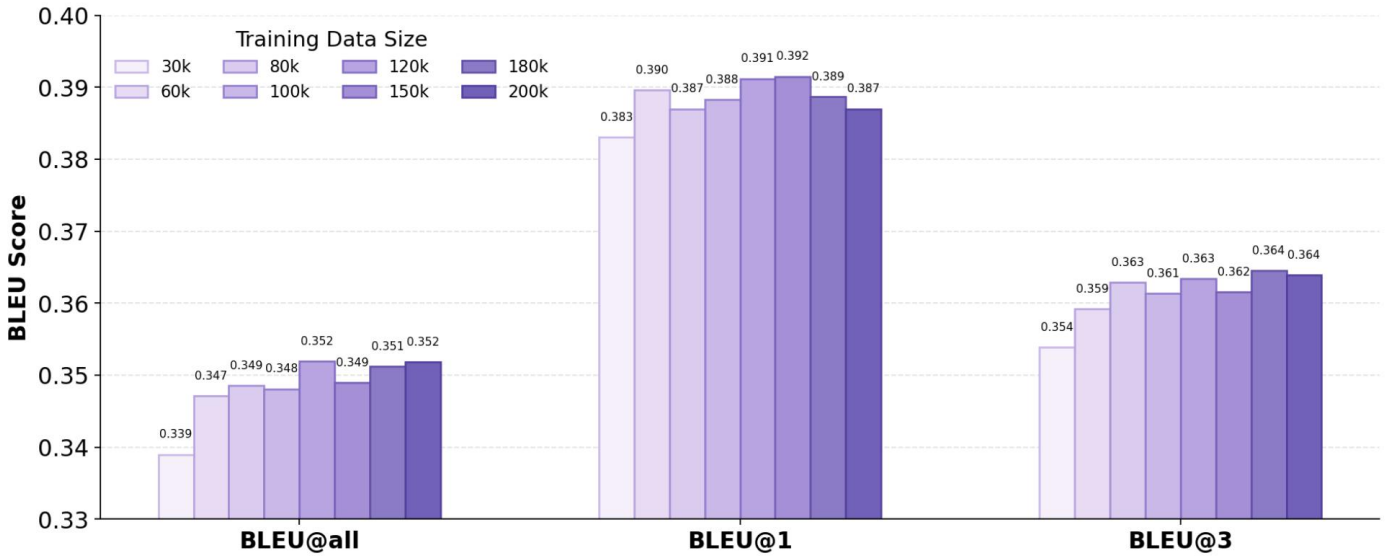
Echo 框架

Model	Codewise-7B-SFT		Codewise-7B-Echo	
Metrics	AR	GR	AR	GR
All	25.01	36.05	37.55	40.58
Top-6 Languages	26.82	37.14	39.70	42.93
Vue	25.22	37.22	38.74	40.15
Python	24.80	41.30	41.78	46.83
Java	28.59	31.84	38.26	38.98
JavaScript/TypeScript	27.29	41.43	40.67	46.46
C/C++	25.30	47.67	39.35	49.03
Go	31.08	37.35	42.91	44.11
Other	20.53	32.96	32.00	34.69

Benefit of Echo

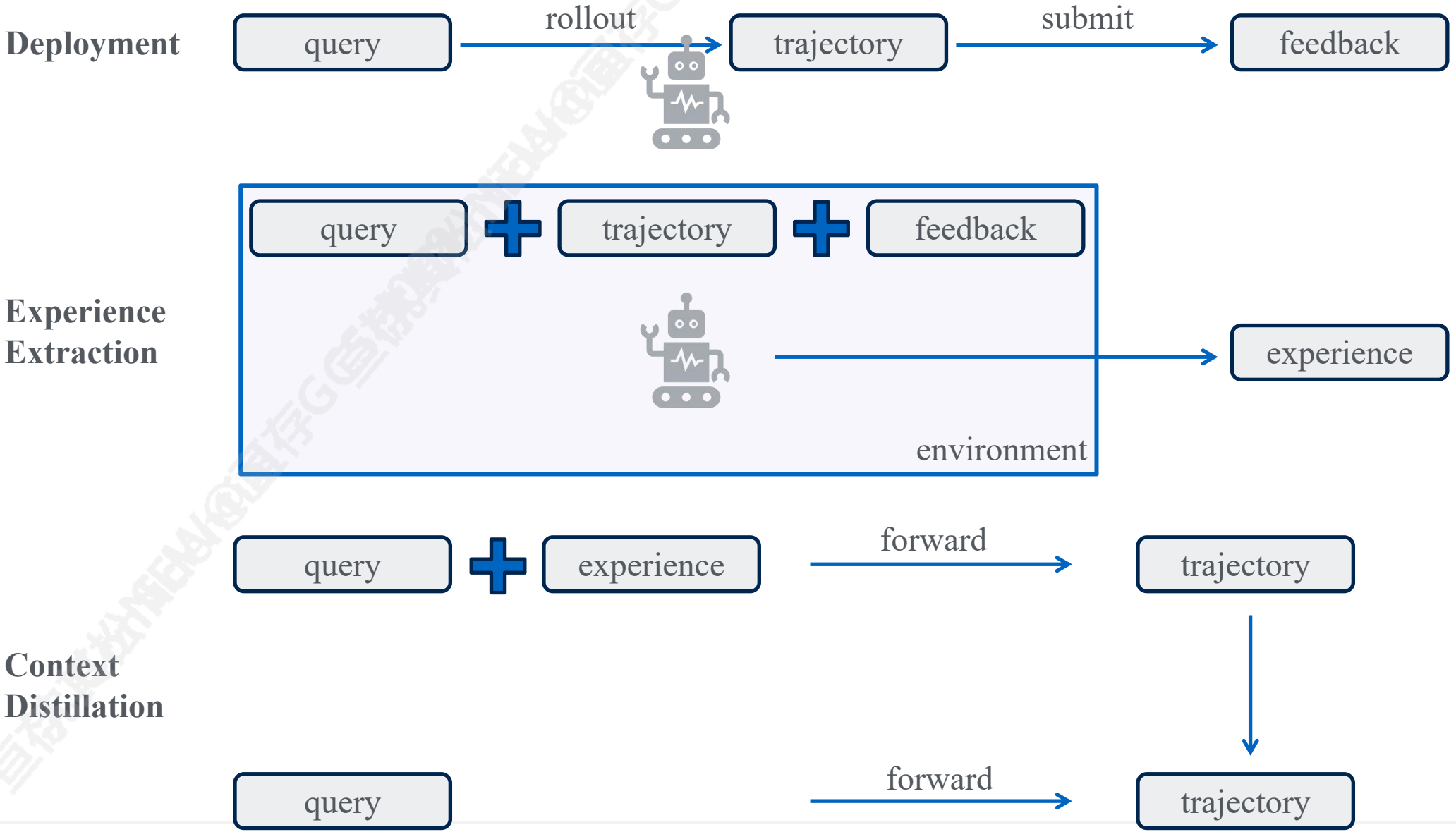


Monthly Acceptance Rate



Scaling Behavior

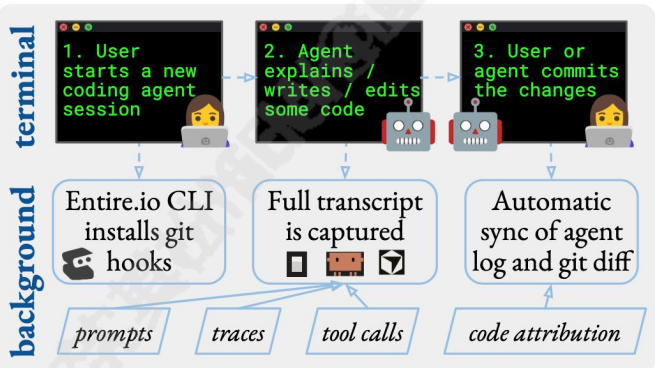
交互经验内化

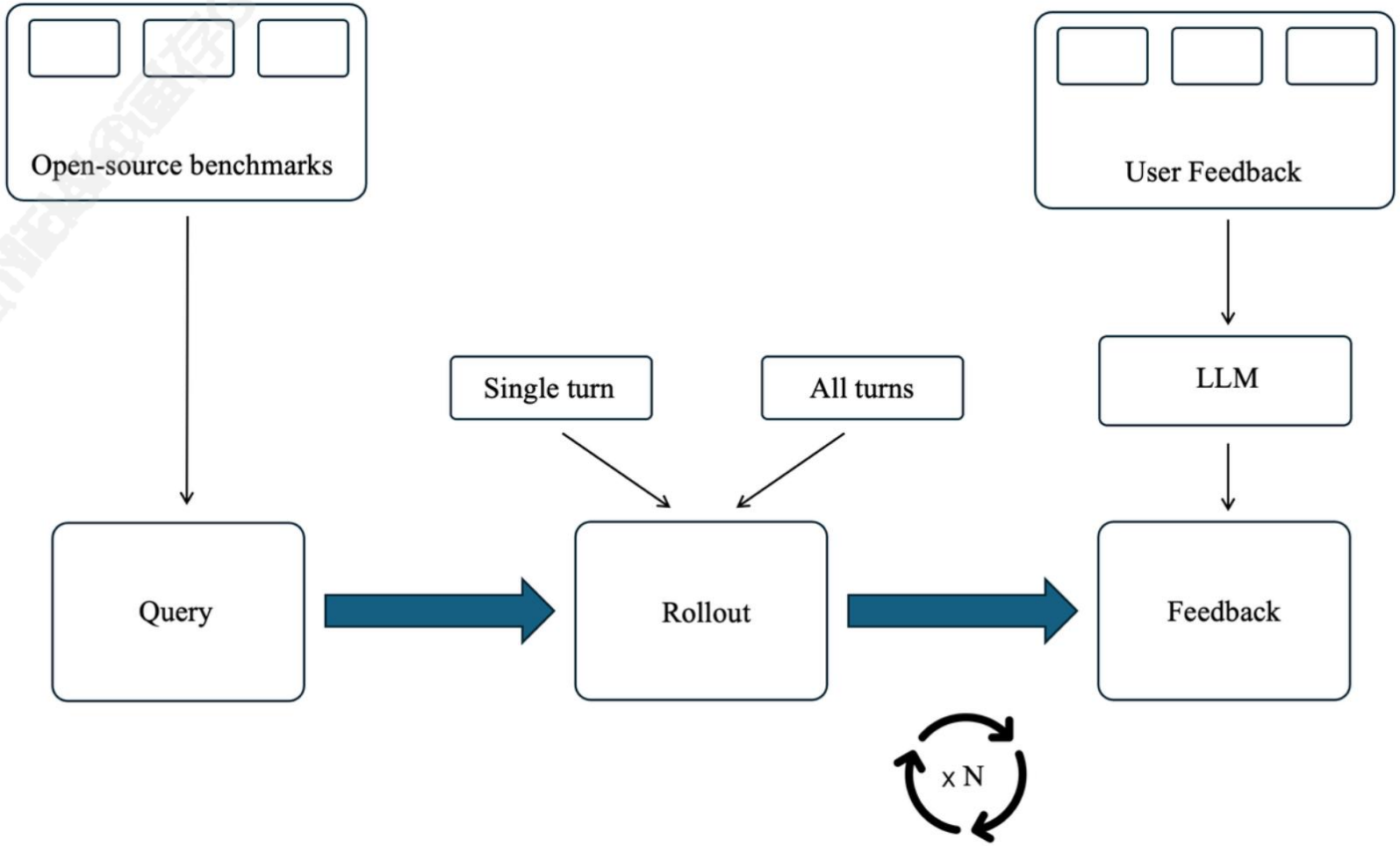

Deployment
query
rollout
trajectory
submit
feedback
Experience Extraction
query
trajectory
feedback
experience
environment
Context Distillation
query
experience
forward
trajectory
query
forward
trajectory

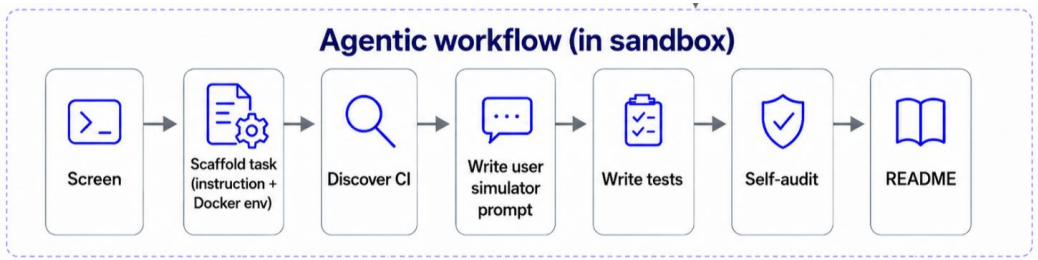
训练集构造

SWE-chat
dataset statistics:

- **205** public GitHub repos
- **6K** coding agent sessions
- **13K** checkpoints
- **63K** user prompts
- **355K** tool calls

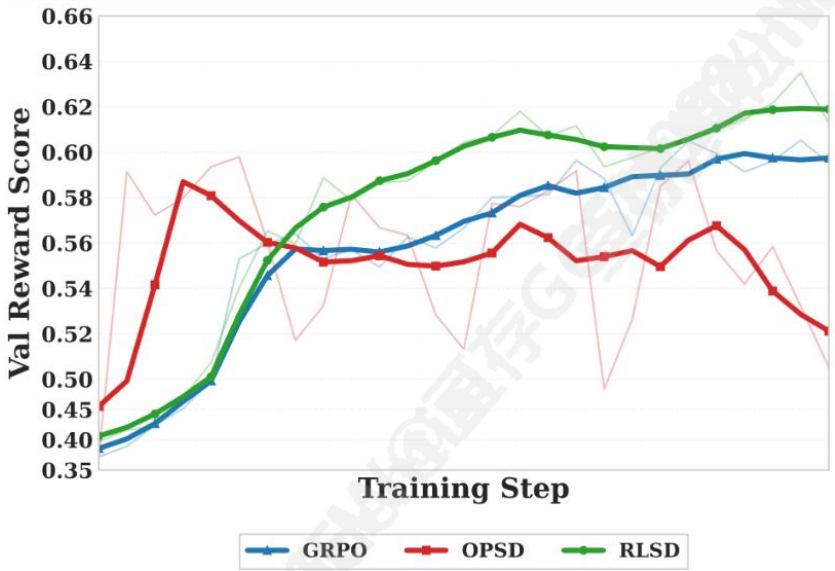


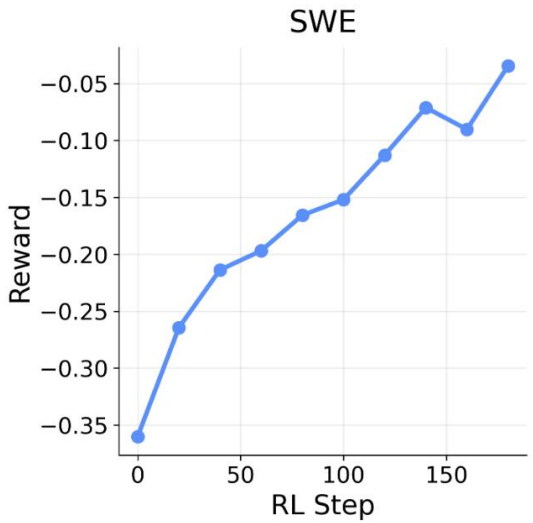




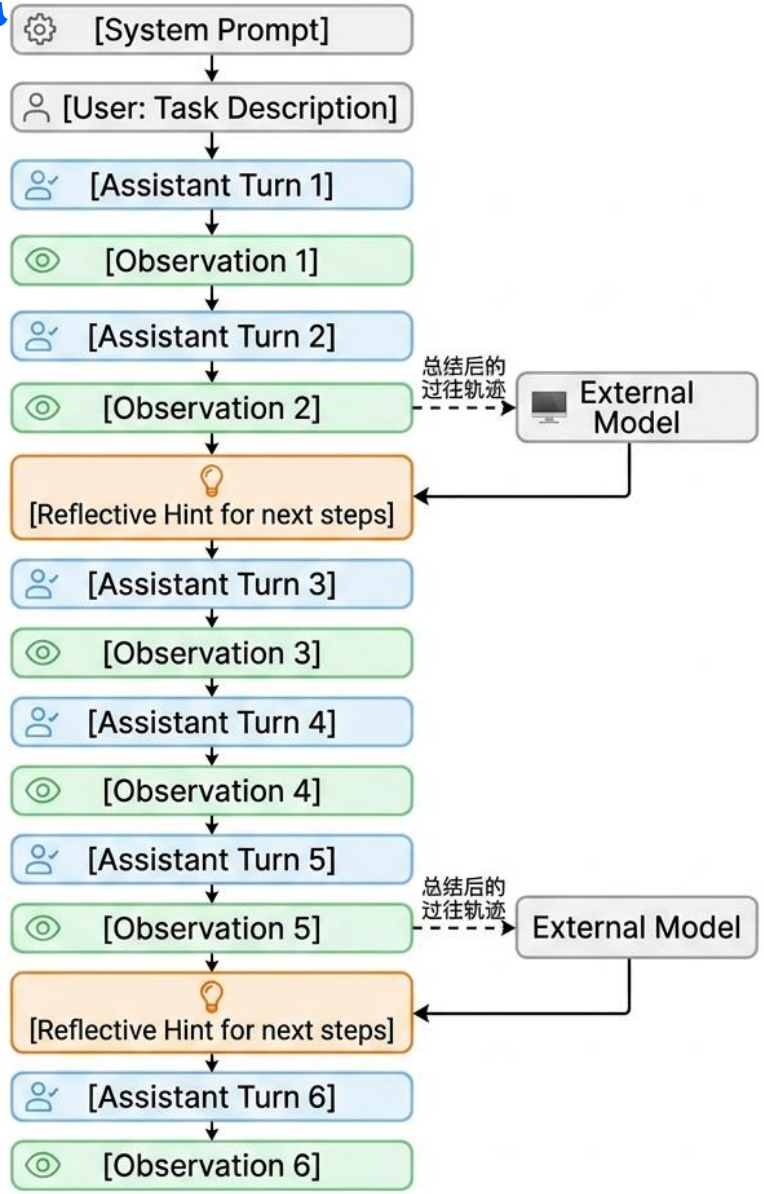
Turn-level & Step-level 经验内化

根据轨迹内的问题步骤，插入Reflective Hint进行OPSD训练
–Cursor Composer 2.5





results	swe-bench verified
Qwen3-30A3B	24%
Qwen3-30A3B + GRPO	30%
Qwen3-30A3B + OPSD	**33%**



总结

Code Agent 如何在训练-测试-部署的闭环中持续提升能力?

01 训练层面

- 样本构建问题
- 训练稳定性问题

02 执行层面

- 执行结果利用问题
- Agent Team协同演进

03 交互层面

- Echo整体范式
- 反馈蒸馏实践

Thanks~

haozhezheng@outlook.com