中国人民大学
RENMIN UNIVERSITY OF CHINA

结构化数据智能：从推理模型到智能体的演进

中国人民大学-金仓联合实验室

张静

2026年11月25日

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

背景与动机

■ **大模型**习得大量人类知识，展现强通用性，但人类希望它适配完成越来越难的任务

问答

直接提供答案



推理

先思考再给答案



智能体

先思考再调用工具



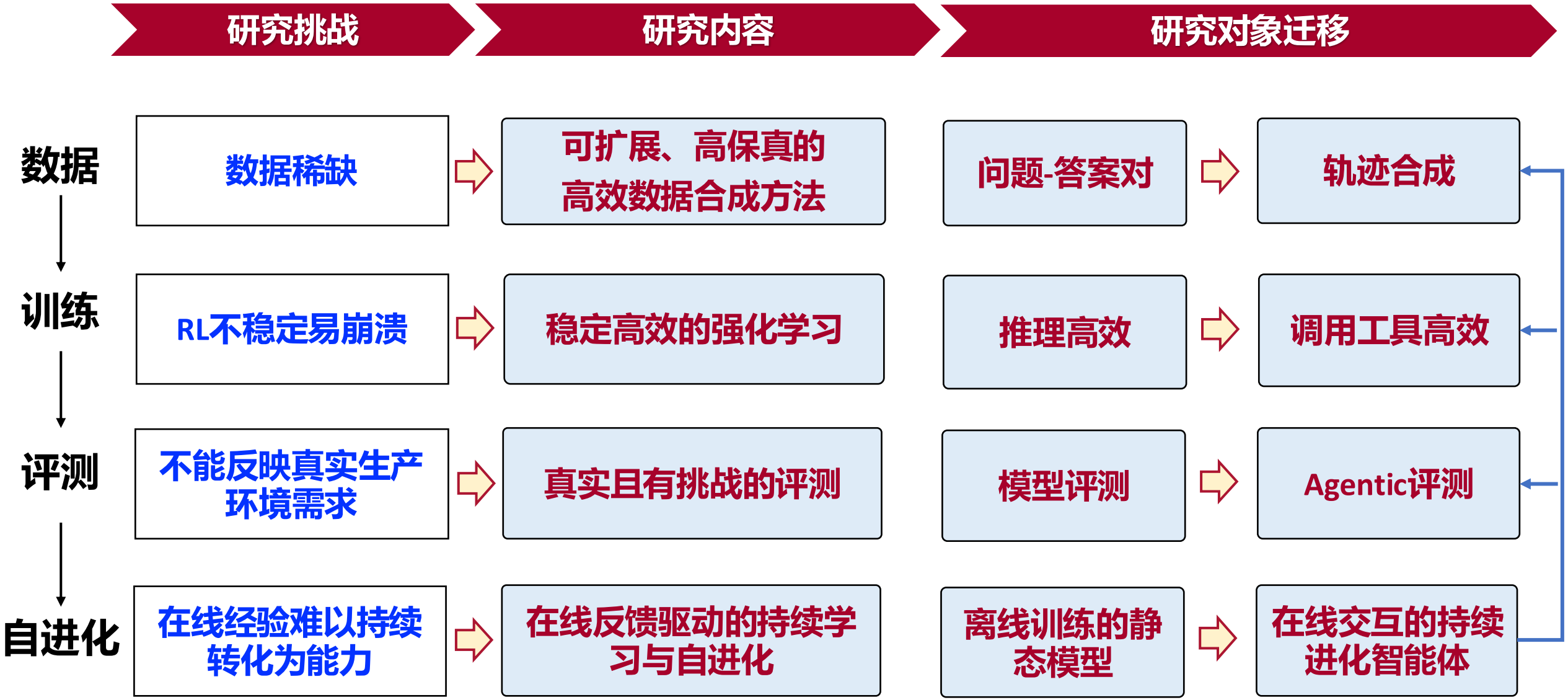
静态回答 → **动态决策**

自主性

长链推理与记忆

环境交互与反馈学习

背景与动机


研究挑战
研究内容
研究对象迁移
数据
训练
评测
自进化
数据稀缺
可扩展、高保真的高效数据合成方法
问题-答案对
轨迹合成
RL不稳定易崩溃
稳定高效的强化学习
推理高效
调用工具高效
不能反映真实生产环境需求
真实且有挑战的评测
模型评测
Agentic评测
在线经验难以持续转化为能力
在线反馈驱动的持续学习与自进化
离线训练的静态模型
在线交互的持续进化智能体

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

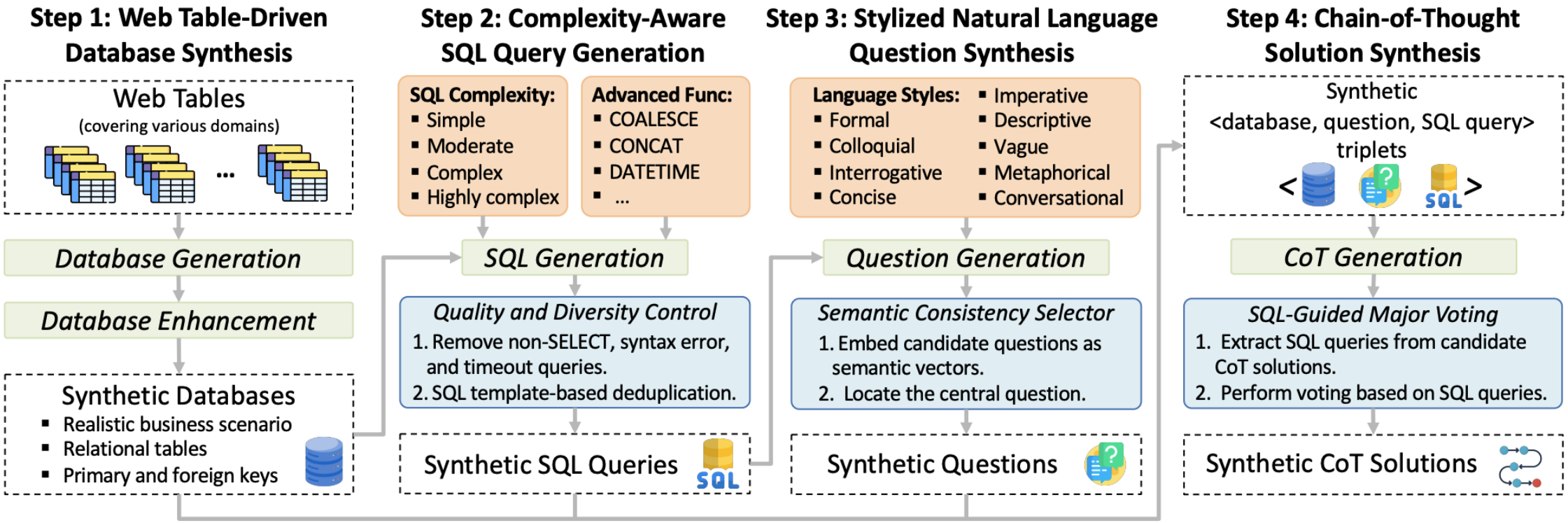
05 技能自进化

06 总结与展望

OmniSQL：可扩展的 Text-to-SQL 数据合成

先建立可执行的数据基础，再扩展问题表达与推理轨迹

Web 表格构造数据库，生成可靠 SQL，再合成多样问题与思维链



Haoyang Li, Shang Wu, Xiaokang Zhang, et al. OmniSQL: Synthesizing High-quality Text-to-SQL Data at Scale.VLDB 25

SynSQL-2.5M：规模、复杂度与合成质量

约 250 万条跨领域 Text-to-SQL 数据，兼顾覆盖和可靠性

Table 2: Overall statistics of different datasets. Note: † denotes that each database in WikiSQL consists of a single table. ∗ indicates that the number of unique SQL queries for BIRD cannot be calculated due to the inaccessibility of its test set.

Dataset	Source	# Example	# Unique SQL	# DB	Lang. Styles	Knowledge	CoT Solution
WIKISQL [85]	Human+Template	80,654	80,257	26,531†	✘	✘	✘
Spider [80]	Human	10,181	4,489	200	✘	✘	✘
BIRD [42]	Human	12,751	-*	95	✘	✓	✘
ScienceBenchmark [81]	LLM-Gen+Human+Template	5,031	3,652	3	✘	✘	✘
EHRSQL [37]	Human+Template	20,108	18,253	2	✘	✘	✘
SynSQL-2.5M	LLM-Gen	2,544,390	2,412,915	16,583	✓	✓	✓

Table 3: Database statistics. The asterisk(*) means BIRD's database statistics are based on training and development sets due to test set inaccessibility.

Dataset	# DB	# Tbl/DB	# Col/DB	# PK/DB	# FK/DB
WIKISQL [85]	26,531	1.00	6.34	1.00	0.00
Spider [80]	200	5.11	26.82	4.70	4.79
BIRD* [42]	95	7.64	54.56	6.71	6.58
ScienceBenchmark [81]	3	16.67	86.67	14.33	20.33
EHRSQL [37]	2	13.50	92.00	13.50	17.00
SynSQL-2.5M	16,583	10.15	74.07	10.14	9.61

百万量级的跨领域训练数据

数据库复杂程度接近真实场景

Table 4: SQL statistics. "Agg." and "Func." are abbreviations of aggregations and functions. The asterisk(*) means BIRD's SQL statistics are based on training and development sets due to test set inaccessibility.

Dataset	# Tables per SQL	# Joins per SQL	# Func. per SQL	# Tokens per SQL	# Agg.	# Set Operators	# Subqueries	# Window Func.	# CTEs	# Unique Skeletons	# Unique Func.
WIKISQL [85]	1.00	0	0.72	10.32	22,707	0	0	0	0	488	5
Spider [80]	1.52	0.48	0.51	14.85	3,802	348	604	0	0	2,136	5
BIRD* [42]	1.98	0.94	0.67	24.75	5,144	25	843	10	9	4,596	24
ScienceBenchmark [81]	1.42	0.42	0.25	14.61	1,234	2	23	0	0	855	5
EHRSQL [37]	3.60	0.62	**2.63**	41.62	11,673	166	17,989	3,207	0	2,447	11
SynSQL-2.5M	**4.00**	**1.75**	1.49	**57.05**	**1,897,440**	**363,472**	**612,647**	**33,916**	**1,073,483**	**2,190,988**	**83**

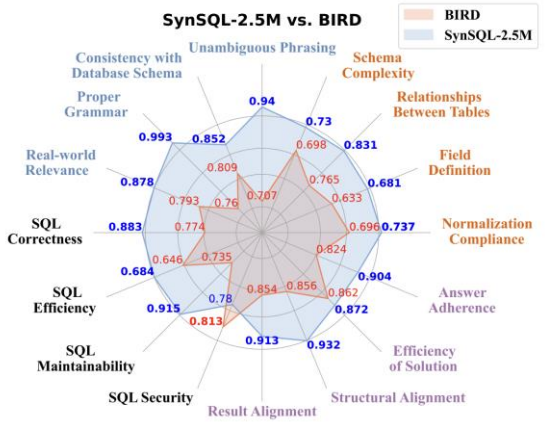


Figure 3: Quality evaluation of SynSQL-2.5M and BIRD.

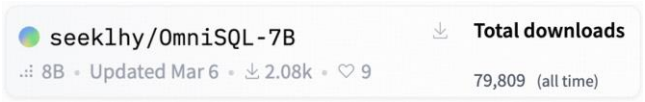
SQL 复杂性与多样性超过已有数据集

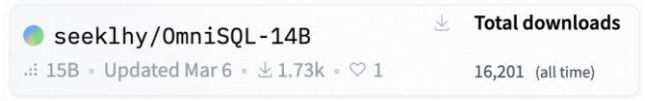
原报告：合成质量评价超过 BIRD 人工标注数据

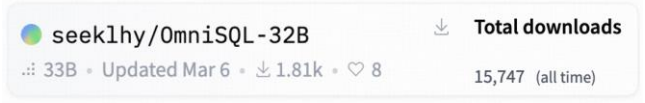
OmniSQL：模型传播与研究影响

原报告统计：Hugging Face 与 ModelScope 累计下载超过 16 万

Hugging Face：11.1 万







ModelScope：5.7 万







基准表现

RESDSQL、CodeS、OmniSQL
在 Spider / BIRD / Spider 2
取得过总计 3 个月榜首

研究影响

AAAI’23 工作被 Google DeepMind 引用，评价为 fine-tuned SOTA

论文影响

原报告统计：SIGMOD’24 工作在当年论文中引用量排名第一

TableLLM：交叉验证提升表格问答保真性

已有问题与答案，仍然需要验证合成思维链是否可靠

示例：Amcor 比 FSD 79 多多少员工?

文本推理路径

读取 Amcor = 350、FSD 79 = 287
计算 350 – 287 = 63
得到文本答案：63

代码执行路径

按 Employer 定位两行记录
提取员工数并执行相减
得到代码答案：63

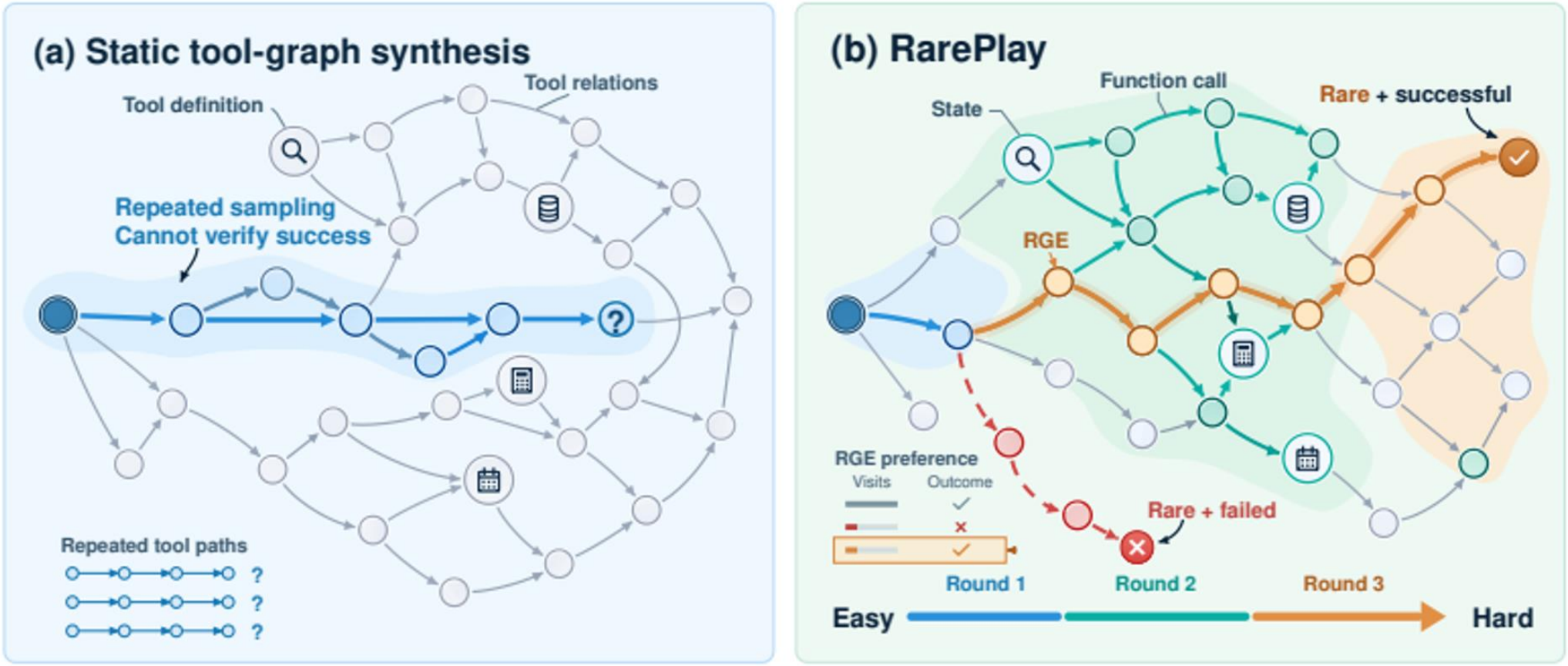
交叉核验：两条路径给出一致答案，保留相应推理样本

理论证明：$P(Y_a) = P(Y_b) = p > 1/2$，满足所列分布假设时 $P(Y|E) \geq p$。
更独立的推理来源有助于减少共同错误，筛选收益依赖相应假设。

Xiaokang Zhang, Sijia Luo, Bohan Zhang, et al. TableLLM: Enabling Tabular Data Manipulation by LLMs in Real Office Usage Scenarios. Findings of ACL 24

RarePlay：状态图让工具任务更可验证

工具图描述调用关系，状态图保留真实环境变化与任务完成依据



状态与行为统计共同定位覆盖缺口，让合成目标跟随当前模型能力

RarePlay: Exploring Scarce Action Spaces in Tool Graphs via Self-Evolving Synthesis. Under review.

RarePlay：RGE 定向采样与迭代合成

采集轨迹建图 → RGE 路径采样 → 逆向合成任务 → 教师执行验真 → 掩码 SFT

$$\mathrm{RGE}(a_i) = r_{a_i} + c_{\mathrm{puct}} \sqrt{\frac{\ln N_{\mathrm{parent}}}{N_i}} + \lambda_{\mathrm{div}} \mathbb{I}[f(a_i) \neq f(a_{i-1})] - \lambda_{\mathrm{trans}} d_i.$$

a_i 为含参数的工具动作；RGE 分数经 Softmax 转成路径采样概率

成功关联

r_{a_i}：含该边的轨迹
最终完成任务的比例

稀缺性奖励

N_i：边的访问次数
N_parent：起点访问次数
c_puct：探索强度

工具切换奖励

$f(a_i)$：本步工具名
I[·]：条件成立取 1
λ_div：切换奖励权重

跨域惩罚

d_i：累计跨业务域次数
λ_trans：对应惩罚权重

更新策略后重新采集轨迹，从头重建状态图

Rubric 检查规划覆盖与结果一致性；教师重新执行合成任务。

RarePlay：行为驱动合成的实验验证

核心贡献：针对当前模型的覆盖缺口，定向合成训练数据

三轮训练，共 4K 轨迹；方法对比匹配骨干与轨迹条数

模型 / 方法	BFCL 多轮均值（%）
Qwen3-4B	21.88
ToolMind-4B	24.25
RarePlay-4B	**30.00**
Qwen3-14B	28.50
RarePlay-14B	**36.00**

消融（4B，BFCL 均值）

设置	准确率（%）
无 RGE	23.38
无拒绝采样	24.75
完整方法	**30.00**

RGE 使合成保留率
从 17.2% 升至 60%

4B 较 ToolMind 高 5.75 个百分点：定向探索与验真共同支持合成收益

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

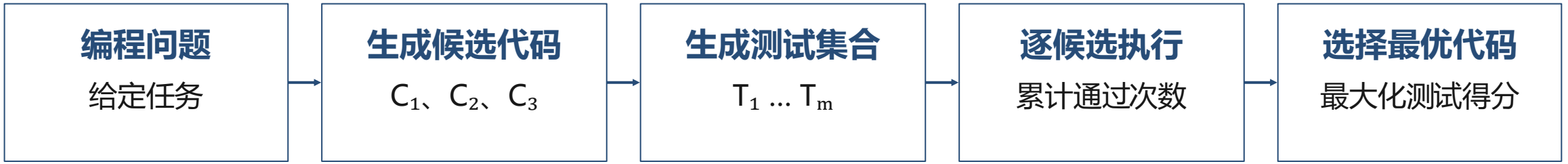
04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

单元测试：用执行结果选择候选代码

生成多个候选程序，用测试通过统计提供代码质量信号



候选代码	通过次数	失败次数
C_1	1	2
C_2	**2**	**1**
C_3	0	3

执行反馈

$r_{i,j} = 1$：C_i 通过测试 T_j

$r_{i,j} = 0$：否则

$$C_opt = \text{argmax}_i \sum_j r_{i,j}$$

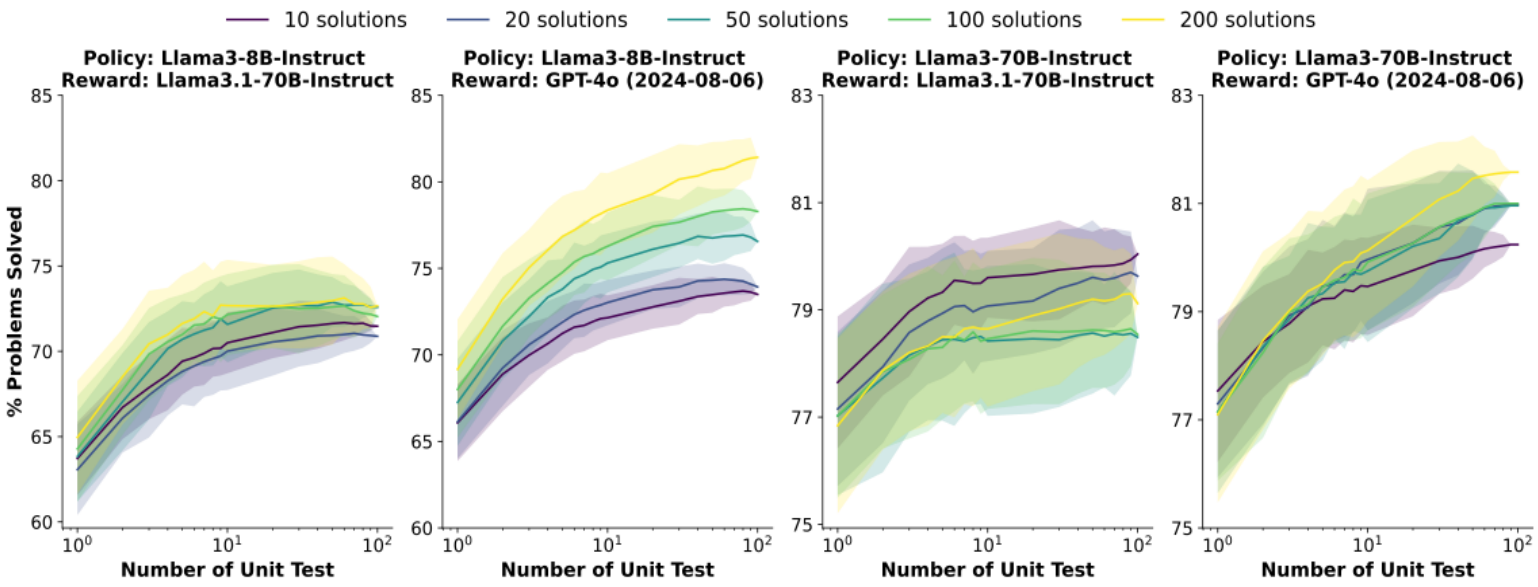
示例选择 C_2，得分来自全部单元测试的累计通过数

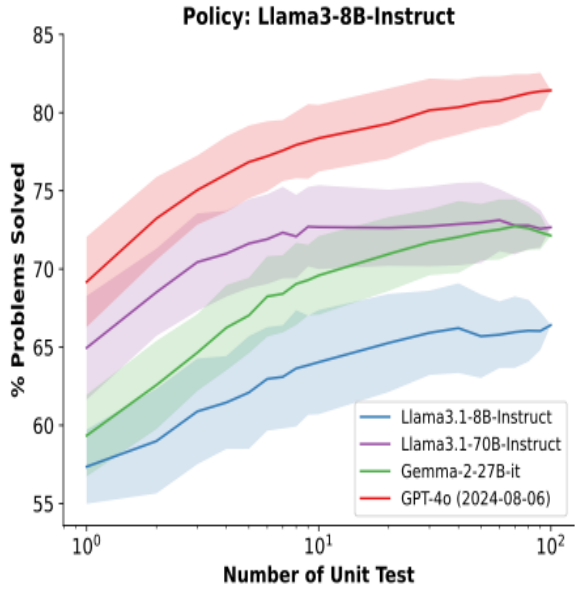
Zeyao Ma, Xiaokang Zhang, Jing Zhang, et al. (2025). Dynamic Scaling of Unit Tests for Code Reward Modeling.
https://arxiv.org/abs/2501.01054

单元测试：规模与生成质量的实验验证

增加测试数量与提升测试质量，能否改善代码筛选?

HumanEval+：策略模型生成候选答案，按单元测试累计通过数选优



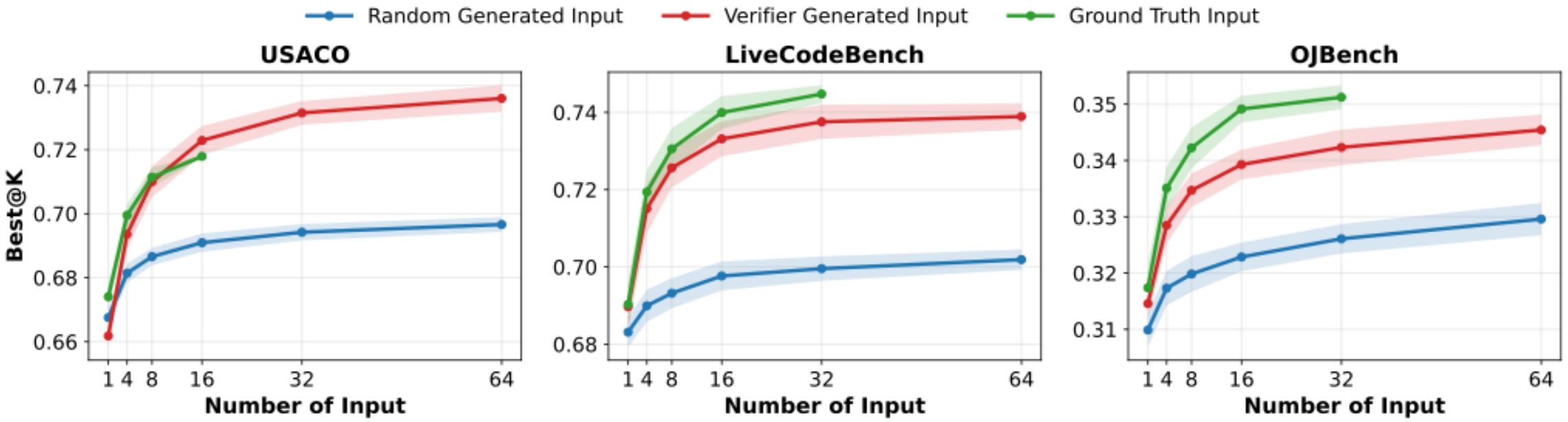


测试规模扩展总体改善筛选，更强的生成模型能更好利用扩展预算

竞赛编程：生成有区分度的测试输入

难题的预期输出难以直接生成，可先寻找区分候选程序的输入

在三个竞赛编程基准上，生成测试输入并执行候选程序，再按输出一致性投票选优

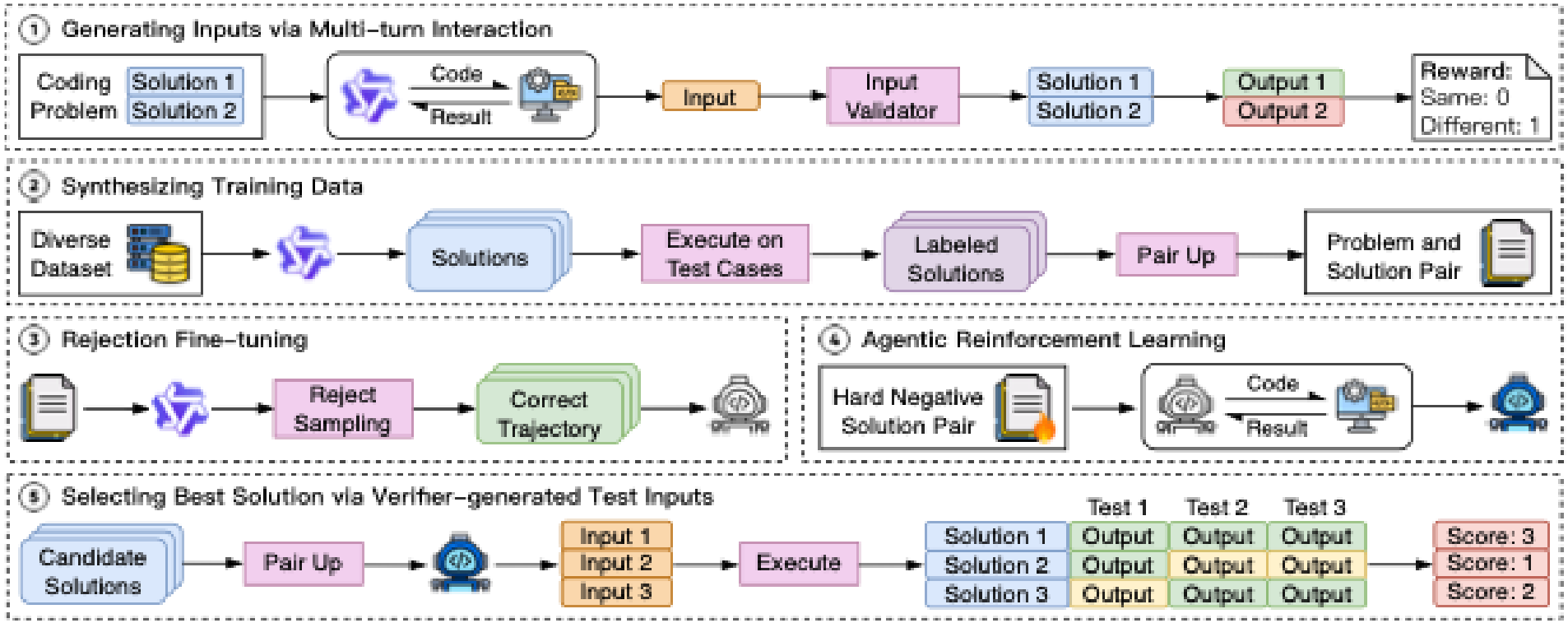


随机输入的扩展收益有限，区分性输入带来更有效的验证

Zeyao Ma, Jing Zhang, Xiaokang Zhang, et al. Scaling Agentic Verifier for Competitive Coding. ICML 26

Agentic Verifier：交互轨迹驱动的验证器训练

通过真实执行反馈学习寻找有效测试输入



交互数据

正确与错误代码配对，
多轮生成并执行测试输入

拒绝式微调

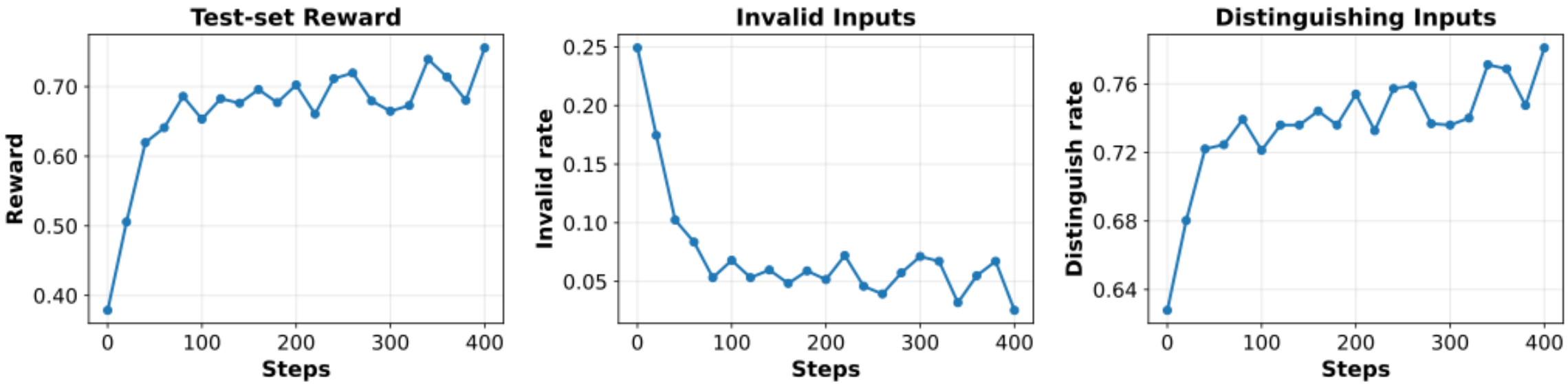
筛选成功区分程序的轨迹，学习基本策略

强化学习

以执行奖励优化
测试输入生成能力

Agentic Verifier：多轮强化学习的执行奖励

把输入合法性与程序区分能力转化为可执行奖励



执行结果	奖励
输入不合法	−1
输入合法，但无法区分候选程序	0
输入合法，且成功区分候选程序	1

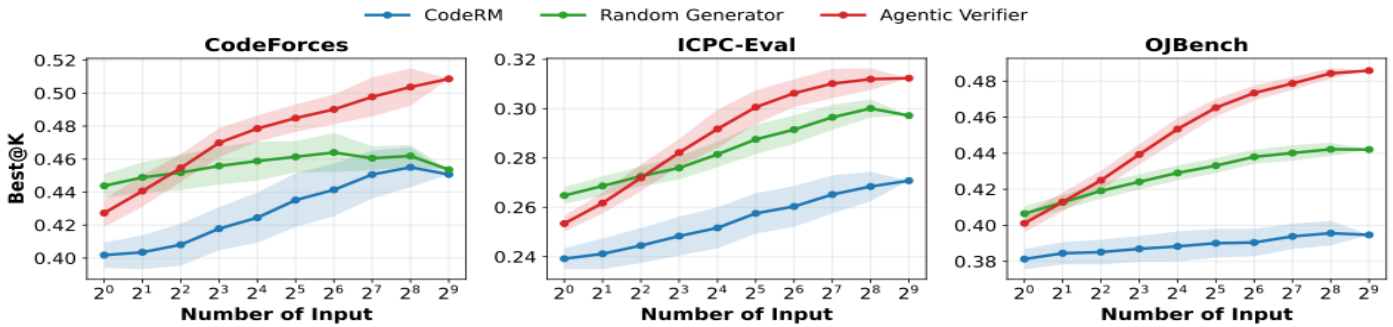
训练趋势

奖励整体提升
不合法输入比例下降
区分性输入比例提高

Agentic Verifier：主实验与测试预算收益

多个竞赛编程基准、两个代码生成模型与多类验证基线

Method / Dataset	USACO		LiveCodeBench		OJBench		ICPC-Eval		CodeForces	
	Best@8	Best@64	Best@8	Best@64	Best@8	Best@64	Best@8	Best@64	Best@8	Best@64
Qwen3-30B-A3B-Thinking-2507										
Vanilla	62.8	62.8	66.2	66.2	28.9	28.9	16.6	16.6	33.2	33.2
Grading RM	63.0 +0.2	61.9 -0.9	67.1 +0.9	68.0 +1.8	28.7 -0.2	29.3 +0.4	18.1 +1.5	13.6 -3.0	33.0 -0.2	32.8 -0.4
MBR-Exec (hard)	64.4 +1.6	68.5 +5.7	68.5 +2.3	69.2 +3.0	29.6 +0.7	32.4 +3.5	17.7 +1.1	19.8 +3.2	33.4 +0.2	33.7 +0.5
MBR-Exec (soft)	67.9 +5.1	69.6 +6.8	70.0 +3.8	70.0 +3.8	31.4 +2.5	31.8 +2.9	20.5 +3.9	20.5 +3.9	38.3 +5.1	41.8 +8.6
CodeT	68.5 +5.7	69.8 +7.0	72.7 +6.5	70.5 +4.3	31.6 +2.7	32.1 +3.2	21.6 +5.0	21.3 +4.7	38.6 +5.4	40.1 +6.9
CodeRM	68.7 +5.9	68.4 +5.6	71.2 +5.0	71.4 +5.2	32.0 +3.1	31.3 +2.4	21.3 +4.7	18.5 +1.9	38.9 +5.7	42.3 +9.1
Random Generator	69.7 +6.9	70.2 +7.4	70.7 +4.5	70.4 +4.2	33.4 +4.5	34.1 +5.2	22.4 +5.8	22.2 +5.6	39.5 +6.3	40.1 +6.9
Agentic Verifier	73.1 +10.3	74.5 +11.7	73.6 +7.4	73.7 +7.5	33.5 +4.6	34.9 +6.0	23.1 +6.5	25.8 +9.2	43.0 +9.8	46.4 +13.2
Qwen3-235B-A22B-Thinking-2507										
Vanilla	75.4	75.4	74.7	74.7	37.1	37.1	23.6	23.6	39.7	39.7
Grading RM	75.8 +0.4	73.6 -1.8	75.9 +1.2	74.9 +0.2	37.6 +0.5	34.9 -2.2	23.5 -0.1	22.0 -1.6	39.7 +0.0	34.4 -5.3
MBR-Exec (hard)	75.8 +0.4	79.2 +3.8	77.9 +3.2	77.7 +3.0	37.7 +0.6	39.6 +2.5	23.9 +0.3	27.4 +3.8	40.0 +0.3	42.2 +2.5
MBR-Exec (soft)	79.9 +4.5	80.3 +4.9	78.5 +3.8	78.5 +3.8	40.5 +3.4	41.4 +4.3	26.2 +2.6	26.4 +2.8	43.8 +4.1	45.4 +5.7
CodeT	79.0 +3.6	80.4 +5.0	78.8 +4.1	79.2 +4.5	41.0 +3.9	40.6 +3.5	26.9 +3.3	27.4 +3.8	44.6 +4.9	46.9 +7.2
CodeRM	78.7 +3.3	77.5 +2.1	79.3 +4.6	80.3 +5.6	40.4 +3.3	39.5 +2.4	26.7 +3.1	27.1 +3.5	45.9 +6.2	45.1 +5.4
Random Generator	81.8 +6.4	82.2 +6.8	78.8 +4.1	78.9 +4.2	43.8 +6.7	44.2 +7.1	28.7 +5.1	29.7 +6.1	46.7 +7.0	45.4 +5.7
Agentic Verifier	83.0 +7.6	84.1 +8.7	81.6 +6.9	82.9 +8.2	45.9 +8.8	48.6 +11.5	29.9 +6.3	31.2 +7.6	49.9 +10.2	50.9 +11.2



实验设置

覆盖多个竞赛编程基准
比较打分与执行验证方法
逐步增加测试输入预算

主要发现

所展示设置中总体表现更好
区分性输入更有效利用
测试计算预算

EG-GRPO：首次证据奖励与知识状态

即使最终答错，前面的搜索也可能很有价值; 即使最终答对，中间仍可能有大量无效操作

合成题自带 K 条原子事实，LLM judge 将最早支持引用映射为轮次 $\varphi_i(c_k)$

步骤奖励：本轮首次新增证据的占比

$$\Delta_{i,t} = \{c_k \in C_q : \phi_i(c_k) = t\},$$

$$r_{i,t} = \frac{|\Delta_{i,t}|}{K} \in [0, 1].$$

知识状态：行动前已获得的事实集合

$$S_{i,t} = \{c_k \in C_q : \phi_i(c_k) < t\}.$$

状态满足 $\varphi_i(c_k) < t$

相同证据集合支持跨轨迹步骤对齐

示例（K = 4）：重复证据不再奖励，状态在行动前统计

搜索轮次	行动前的状态	本轮首次新增证据	步骤奖励
1	∅	$\{c_1\}$	1 / 4 = 0.25
2	$\{c_1\}$	∅	0
3	$\{c_1\}$	$\{c_2, c_3\}$	2 / 4 = 0.50
4	$\{c_1, c_2, c_3\}$	$\{c_4\}$	1 / 4 = 0.25

事实清单仅用于训练时监督，测试问题无需附带事实清单

Sijia Luo, Mengjie Gao, Yuxiao He, et al. (2026). EG-GRPO: Evidence-Guided Step-level Credit Assignment for Deep Search Agents. https://arxiv.org/abs/2610.25269

EG-GRPO：知识状态分组与优势估计

在相近证据进度下，比较后续搜证回报

$$G_{i,t} = \sum_{k=t}^{T} \gamma^{k-t} r_{i,k},$$

γ = 0.95：后续证据收益向前分配信用
支持有价值的铺垫步骤，仍非因果归因

层级 / 比较对象	分组键	样本阈值	含义
1 同题、相同事实集合	$(q, S_{i,t})$	≥ M_1 个步骤	条件最接近，样本可能稀疏
2 同题、相同事实数量	$(q, \lvert S_{i,t}\rvert)$	≥ M_1 个步骤	放宽事实内容，增加样本
3 同题的全部有效步骤	q	≥ 2 个步骤	最粗分组，保证比较基线

按顺序选择第一个满足阈值的组，记为 $\mathcal{G}_{i,t}$

选中组内归一化：步骤优势

$$A^S_{i,t} = \frac{G_{i,t} - \text{mean}\left(\{G_{j,u} : (j,u) \in \mathcal{G}_{i,t}\}\right)}{\text{std}\left(\{G_{j,u} : (j,u) \in \mathcal{G}_{i,t}\}\right)}.$$

与最终结果优势融合

$$A_{i,t} = A^T_i + w \cdot \text{clip}\left(A^S_{i,t}, -C, C\right),$$

A^T 为同题结果优势，w = 0.5、C = 5，只更新决策 token

EG-GRPO：过程证据与结果监督的互补

相较 GRPO，五项均值：4B 提升 4.96，30B 提升 3.64 个百分点

4B / 30B Thinking，1K 轨迹 SFT 冷启动，再以 2,316 个合成问题进行 RL

模型 / 方法	BrowseComp	BC-ZH	GAIA	xB-2505	xB-2510
4B + GRPO	6.7	16.7	34.8	62.7	15.7
4B + STAMP	7.3	21.5	35.2	60.0	17.0
4B + EG-GRPO	**8.0**	**23.5**	**45.9**	**65.0**	**19.0**
30B + GRPO	10.3	22.4	42.7	57.6	20.1
30B + STAMP	10.8	24.1	45.2	62.5	21.1
30B + EG-GRPO	**11.1**	**26.1**	**49.1**	**64.2**	**20.8**

核心贡献：证据用于计算奖励，也用于对齐状态、比较未来回报

消融：适量步骤优势与三级回退有效，过大的步骤权重会削弱收益。

Sparse-RL：长轨迹强化学习的效率与稳定性

长序列 rollout 的 KV Cache 形成显存瓶颈

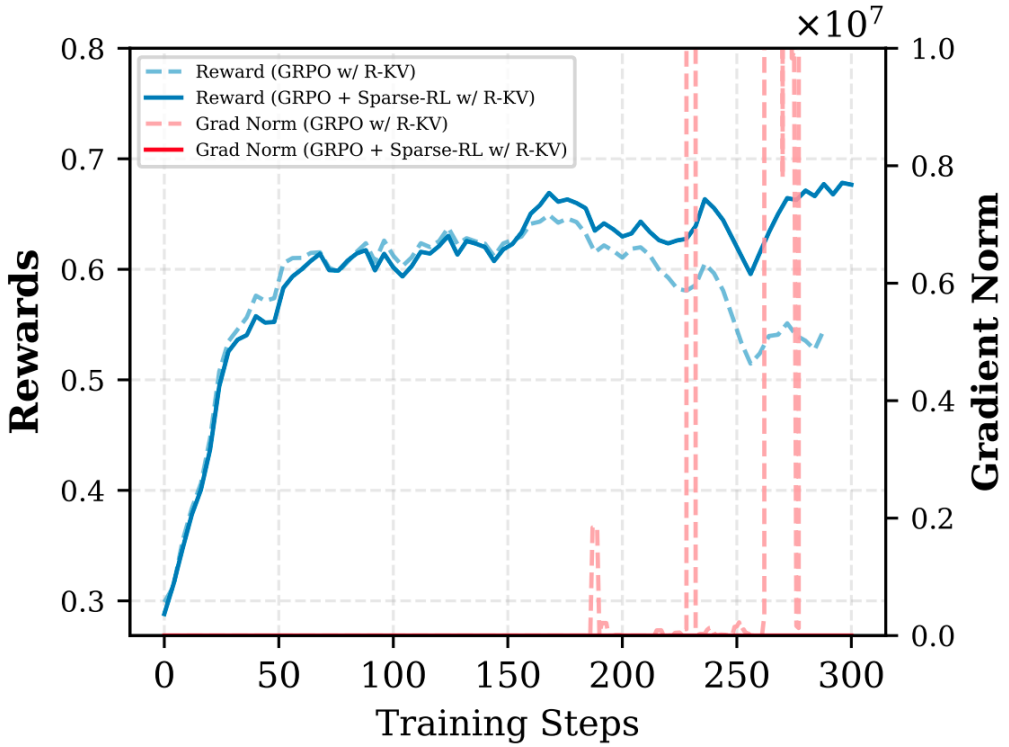
Rollout 主导训练时间，缓存压缩却会改变实际采样行为

策略不一致

压缩 KV 的稀疏策略负责采样，
参数更新仍使用稠密策略，
采样分布与学习分布存在偏差。

异常轨迹

信息丢失可能触发重复序列，
放大负梯度与训练噪声。



直接套用缓存压缩可能导致训练崩溃，需要显式处理策略差异

Sijia Luo, Xiaokang Zhang, Yuxuan Hu, et al. . Sparse-RL: Breaking the Memory Wall in LLM Reinforcement Learning via Stable Sparse Rollouts. ACL 26.

Sparse-RL：分解策略差异并修正训练偏差

区分旧稠密策略、稀疏采样策略与当前学习策略

总体不一致可分解为：策略滞后性与稀疏引入的不一致

$$\frac{\pi_\theta(o_t|x,o_{<t})}{\pi_{\theta_{\text{sparse}}}(o_t|x,o_{<t})} = \underbrace{\frac{\pi_\theta(o_t|x,o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|x,o_{<t})}}_{\text{policy staleness}} \times \underbrace{\frac{\pi_{\theta_{\text{old}}}(o_t|x,o_{<t})}{\pi_{\theta_{\text{sparse}}}(o_t|x,o_{<t})}}_{\text{sparsity-induced mismatch}}$$

异常样本筛选

利用稀疏 / 稠密策略的 token 概率比率，识别并过滤不可靠序列

$$\xi_t = \frac{\pi_{\theta_{\text{old}}}(o_t|x,o_{<t})}{\pi_{\theta_{\text{sparse}}}(o_t|x,o_{<t})} \qquad M^{\text{RS}}(o) = \begin{cases} 0 & \text{if } \exists t \in [1,|o|] \text{ s.t. } \xi_t < \epsilon \\ 1 & \text{otherwise} \end{cases}$$

有效样本重加权

用概率比率修正策略梯度，缓解有效样本中的分布偏差

$$\mathcal{J}_{\text{Sparse-RL}}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D} \\ o_i\sim\pi_{\theta_{\text{sparse}}}}} \left[\frac{1}{G}\sum_{i=1}^{G} M^{\text{RS}}(o_i) \cdot \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \xi_{i,t} \cdot \min\left(w_{i,t}(\theta)\hat{A}_i, \text{clip}\left(w_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i \right) \right]$$

Sparse-RL：性能、训练行为与稀疏推理鲁棒性

效率优化同时检查能力保持、训练稳定性与最终推理行为

Model	Rollout Method	GSM8K	MATH500	Gaokao	Minerva Math	Olympiad	AIME24	AMC23	Avg.	Toks. saving
Llama-3.2-1B-Instruct										
Base	–	36.2	22.8	17.9	5.5	5.5	1.2	8.8	14.0	-
GRPO	Dense	51.2	33.6	**28.8**	**8.8**	8.7	2.9	12.8	21.0	-
GRPO	w/ R-KV	49.8	30.0	23.9	5.5	7.6	1.7	12.8	18.8	35.1%
↪ + Sparse-RL		49.1	30.2	26.5	5.5	8.3	2.1	**15.2**	19.6	
GRPO	w/ SnapKV	52.2	33.6	26.5	5.5	9.0	2.9	13.8	20.5	
↪ + Sparse-RL		**54.6**	**34.4**	27.0	6.2	**9.3**	**3.0**	13.6	**21.2**	
Qwen2.5-1.5B										
Base	–	43.5	21.0	18.4	4.0	4.0	0.3	7.5	14.1	-
GRPO	Dense	72.9	**58.0**	46.5	16.9	21.6	**3.6**	28.5	35.4	-
GRPO	w/ R-KV	48.8	34.0	32.5	7.0	11.6	1.0	21.2	22.3	53.3%
↪ + Sparse-RL		**75.3**	55.8	48.1	**20.2**	**21.9**	2.9	29.5	**36.2**	
GRPO	w/ SnapKV	66.3	37.6	33.8	8.1	16.1	3.1	18.0	26.1	
↪ + Sparse-RL		73.7	57.6	**48.8**	18.0	20.7	3.4	29.5	36.0	
Qwen2.5-3B										
Base	–	76.0	55.8	45.2	19.9	21.6	4.1	27.3	35.7	-
GRPO	Dense	83.5	64.2	**56.9**	25.0	**29.8**	**6.0**	38.6	**43.4**	-
GRPO	w/ R-KV	79.0	56.6	46.2	21.7	23.0	4.7	27.0	36.9	42.0%
↪ + Sparse-RL		**84.6**	**65.6**	53.5	**26.5**	27.7	4.6	37.4	42.8	
GRPO	w/ SnapKV	79.0	54.2	48.3	19.5	23.6	5.2	30.5	37.2	
↪ + Sparse-RL		83.4	64.0	54.8	25.4	25.9	5.3	**39.1**	42.6	
Qwen2.5-7B										
Base	–	81.6	57.4	52.2	22.1	26.4	7.3	35.5	40.4	-
GRPO	Dense	91.0	**73.8**	**63.9**	**37.1**	**39.1**	**14.5**	52.5	**53.1**	-
GRPO	w/ R-KV	88.0	64.2	57.1	23.2	30.7	8.8	43.8	45.1	39.4%
↪ + Sparse-RL		**91.5**	72.0	61.8	32.7	36.6	12.5	**52.7**	51.4	
GRPO	w/ SnapKV	73.4	54.6	46.0	16.9	19.0	2.6	25.2	34.0	
↪ + Sparse-RL		90.1	71.4	60.8	26.8	32.9	10.2	47.6	48.5	

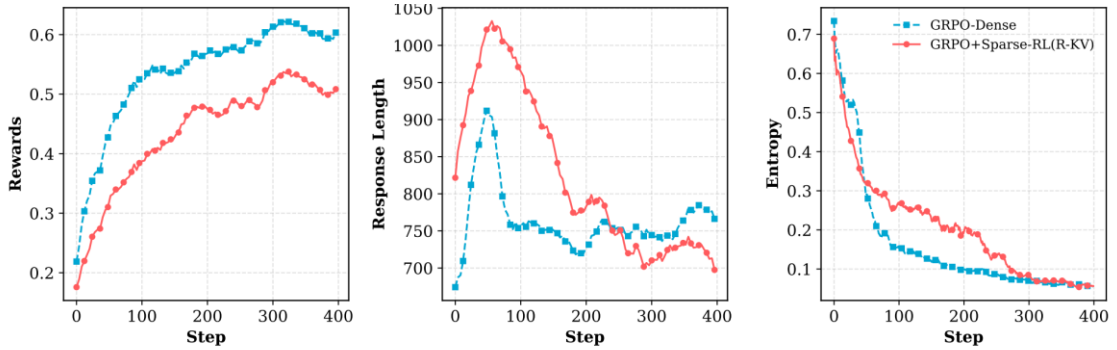
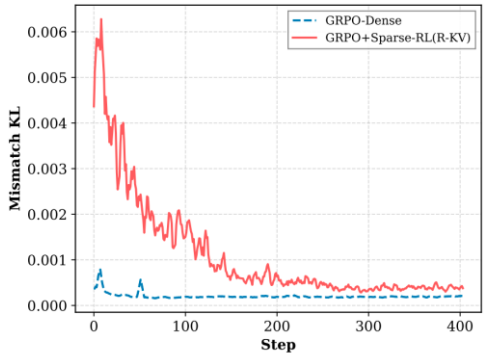


Figure 2: Comparison of training curves between GRPO-Dense and GRPO + Sparse-RL (with R-KV) on Qwen2.5-3B over 400 training steps. **Left**: Average rewards per step. **Middle**: Average response length (tokens). **Right**: Policy entropy.

训练曲线



策略不一致程度

Model	GSM8K	MATH500	Gaokao	Minerva	Olympiad
Qwen2.5-1.5B					
GRPO (Dense)	73.4	50.8	37.7	12.5	14.2
↪ + Sparse-RL (R-KV)	**74.6**	**54.0**	**42.9**	**15.4**	**18.1**
Qwen2.5-3B					
GRPO (Dense)	**85.1**	54.2	44.2	16.5	17.3
↪ + Sparse-RL (R-KV)	84.1	**61.8**	**49.1**	**22.1**	**23.3**
Qwen2.5-7B					
GRPO (Dense)	91.2	71.6	55.3	**33.8**	33.0
↪ + Sparse-RL (R-KV)	**91.4**	**71.8**	**59.0**	30.9	**33.5**

Table 2: The performance of models trained by GRPO + Sparse-RL (with R-KV) and GRPO-Dense in the sparse inference scenario by R-KV compression.

稀疏推理鲁棒性

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

SpreadsheetBench：真实操作与完整业务工作流

评测单元从局部表格操作，扩展到可交付的专业工作簿

维度	SpreadsheetBench 1	SpreadsheetBench 2
任务来源	912 条真实 Excel 论坛指令	321 个专家整理的商业任务
文件复杂度	非标准表格、多表结构与复杂格式	平均 11.8 个工作表 平均 593.5 个单元格修改
能力范围	基于真实用户需求完成表格操作	财务模型 / 模板生成 系统调试 / 可视化
验证重点	每条指令 3 个测试用例 检查操作逻辑能否迁移	跨表与跨步骤依赖 最终工作簿的正确性与完整性

真实业务评测需要检查“能否完整交付”，并覆盖操作逻辑与依赖关系

Zeyao Ma, Bohan Zhang, Jing Zhang, et al. (2024). SpreadsheetBench: Towards Challenging Real World Spreadsheet Manipulation. NeurIPS 24
Jian Zhu, Yuzheng Zhang, Zeyao Ma, et al. (2026). SpreadsheetBench 2: Evaluating Agents on End-to-End Business Spreadsheet Workflows. NeurIPS 26

SpreadsheetBench 2：专业工作流仍有明显挑战

大多数模型整体得分低于 25%，最佳为 34.89%

Model	Template		Financial Model		Debugging		Visualization	Overall	
	Modif.	Acc.	Modif.	Acc.	Modif.	Acc.	Acc.	Modif.	Acc.
Closed-source Models									
Claude Opus 4.6	**91.97**	**52.58**	**89.69**	**34.00**	**50.38**	**12.00**	**62.50**	**77.20**	**34.89**
GPT-5.2	83.84	35.05	87.07	33.00	39.07	8.00	45.83	69.85	26.79
Gemini 3.1 Pro	86.06	28.87	85.26	31.00	41.86	7.00	41.67	70.91	23.68
Open-source Models									
GLM-5	66.60	17.53	78.44	22.00	29.06	7.00	37.50	57.95	17.14
Deepseek-V3.2	75.19	25.77	59.78	7.00	28.71	10.00	33.33	54.35	15.58
Kimi K2.5	67.08	18.56	64.86	15.00	19.48	4.00	41.67	50.31	14.64
Qwen3.5-397B-A17B	65.62	17.53	69.59	10.00	24.06	3.00	25.00	52.96	11.22
MiniMax M2.5	50.14	7.22	62.05	8.00	16.76	4.00	16.67	42.91	7.17

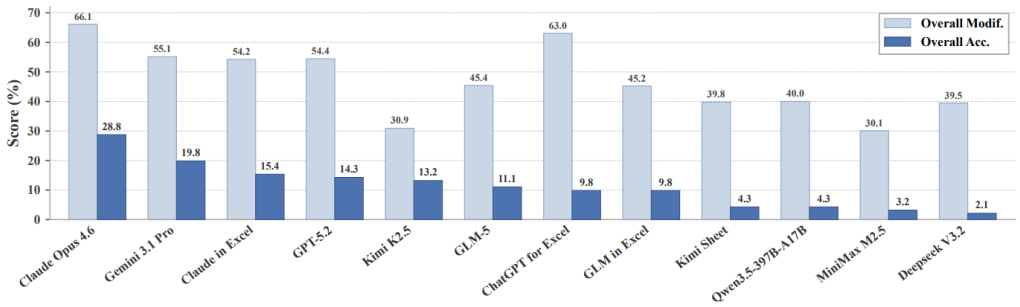


Figure 3: Performance on a 30-example representative subset covering Financial Modeling, Debugging, and Visualization tasks. We compare different models and four LLM-based spreadsheet tools. The results show that LLM-based spreadsheet tools do not outperform models under our agent scaffold, with Claude in Excel achieving the best performance among the spreadsheet tools.

实验设置

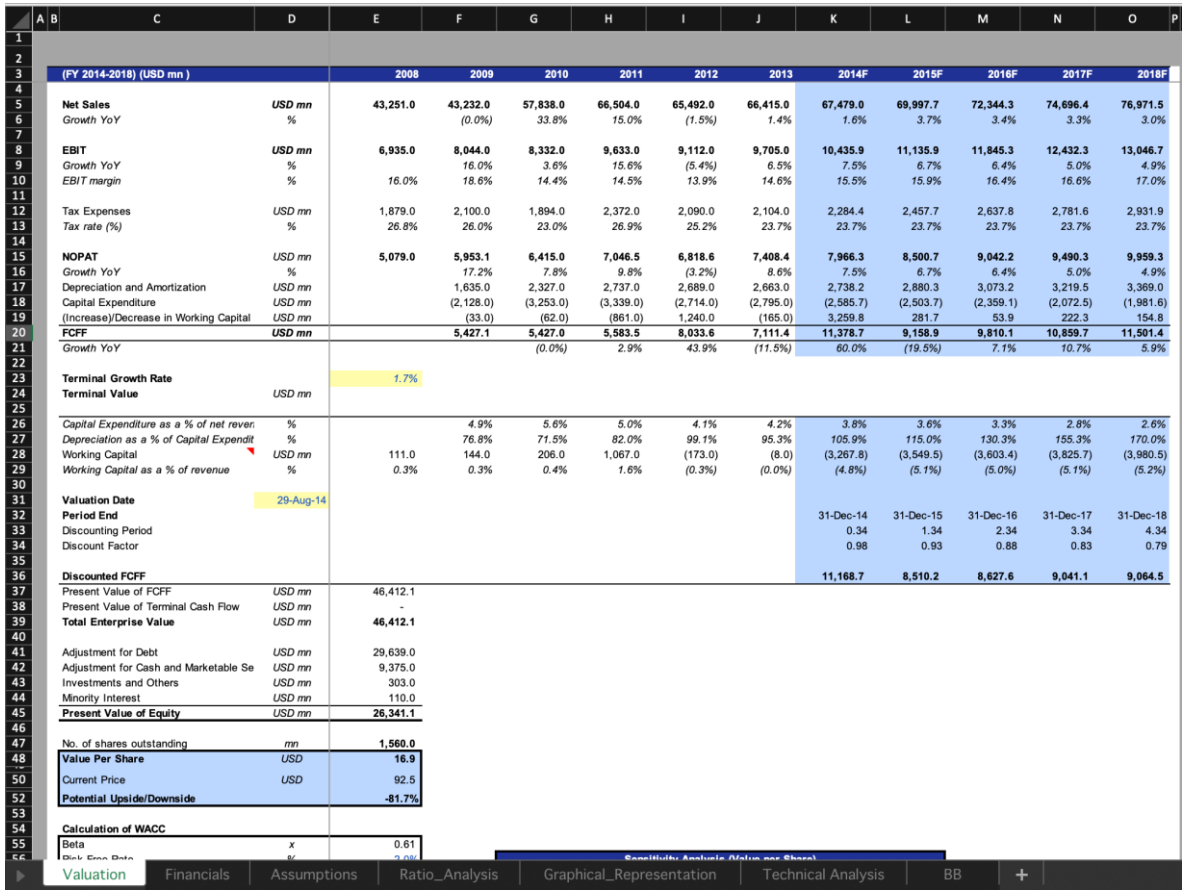
统一 SWE-Agent 框架，使用 view、submit 与 bash 工具执行任务

任务差异：Visualization 相对容易，Financial Modeling 居中，Debugging 最难

财务建模示例：跨工作表依赖与交付约束

完成业务计算，同时保留工作簿已有结构、版式与格式

工作表	需要完成的计算
Financials	历史期与预测期的 CAGR：Revenue、EBITDA、EBIT、PAT、Total Assets
Assumptions	前两年平均值推导 2014F–2018F Average Days Inventory；2015F–2018F 每年减少 0.5 天，再推导预测存货
Ratio Analysis	Total Asset Turnover 与 EV/EBITDA
Valuation	Terminal Value（终值）



(FY 2014-2018) (USD mn)		2008	2009	2010	2011	2012	2013	2014F	2015F	2016F	2017F	2018F
Net Sales	*USD mn*	43,251.0	43,232.0	57,838.0	66,504.0	65,492.0	66,415.0	67,479.0	69,997.7	72,344.3	74,696.4	76,971.5
Growth YoY	%		*(0.0%)*	*33.8%*	*15.0%*	*(1.5%)*	*1.4%*	*1.6%*	*3.7%*	*3.4%*	*3.3%*	*3.0%*
EBIT	*USD mn*	6,935.0	8,044.0	8,332.0	9,633.0	9,112.0	9,705.0	10,435.9	11,135.9	11,845.3	12,432.3	13,046.7
Growth YoY	%		*16.0%*	*3.6%*	*15.6%*	*(5.4%)*	*6.5%*	*7.5%*	*6.7%*	*6.4%*	*5.0%*	*4.9%*
EBIT margin	%	*16.0%*	*18.6%*	*14.4%*	*14.5%*	*13.9%*	*14.6%*	*15.5%*	*15.9%*	*16.4%*	*16.6%*	*17.0%*
Tax Expenses	*USD mn*	1,879.0	2,100.0	1,894.0	2,372.0	2,090.0	2,104.0	2,284.4	2,457.7	2,637.8	2,781.6	2,931.9
Tax rate (%)	%	*26.8%*	*26.0%*	*23.0%*	*26.9%*	*25.2%*	*23.7%*	*23.7%*	*23.7%*	*23.7%*	*23.7%*	*23.7%*
NOPAT	*USD mn*	5,079.0	5,953.1	6,415.0	7,046.5	6,818.6	7,408.4	7,966.3	8,500.7	9,042.2	9,490.3	9,959.3
Growth YoY	%		*17.2%*	*7.8%*	*9.8%*	*(3.2%)*	*8.6%*	*7.5%*	*6.7%*	*6.4%*	*5.0%*	*4.9%*
Depreciation and Amortization	*USD mn*		1,635.0	2,327.0	2,737.0	2,689.0	2,663.0	2,738.2	2,880.3	3,073.2	3,219.5	3,369.0
Capital Expenditure	*USD mn*		(2,128.0)	(3,253.0)	(3,339.0)	(2,714.0)	(2,795.0)	(2,585.7)	(2,503.7)	(2,359.1)	(2,072.5)	(1,981.6)
(Increase)/Decrease in Working Capital	*USD mn*		(33.0)	(62.0)	(861.0)	1,240.0	(165.0)	3,259.8	281.7	53.9	222.3	154.8
FCFF	***USD mn***		**5,427.1**	**5,427.0**	**5,583.5**	**8,033.6**	**7,111.4**	**11,378.7**	**9,158.9**	**9,810.1**	**10,859.7**	**11,501.4**
Growth YoY				*(0.0%)*	*2.9%*	*43.9%*	*(11.5%)*	*60.0%*	*(19.5%)*	*7.1%*	*10.7%*	*5.9%*
Terminal Growth Rate		1.7%										
Terminal Value	*USD mn*											
Capital Expenditure as a % of net rever	%		*4.9%*	*5.6%*	*5.0%*	*4.1%*	*4.2%*	*3.8%*	*3.6%*	*3.3%*	*2.8%*	*2.6%*
Depreciation as a % of Capital Expendit	%		*76.8%*	*71.5%*	*82.0%*	*99.1%*	*95.3%*	*105.9%*	*115.0%*	*130.3%*	*155.3%*	*170.0%*
Working Capital	*USD mn*	111.0	144.0	206.0	1,067.0	(173.0)	(8.0)	(3,267.8)	(3,549.5)	(3,603.4)	(3,825.7)	(3,980.5)
Working Capital as a % of revenue	%	*0.3%*	*0.3%*	*0.4%*	*1.6%*	*(0.3%)*	*(0.0%)*	*(4.8%)*	*(5.1%)*	*(5.0%)*	*(5.1%)*	*(5.2%)*
Valuation Date	29-Aug-14											
Period End								31-Dec-14	31-Dec-15	31-Dec-16	31-Dec-17	31-Dec-18
Discounting Period								0.34	1.34	2.34	3.34	4.34
Discount Factor								0.98	0.93	0.88	0.83	0.79
Discounted FCFF								**11,168.7**	**8,510.2**	**8,627.6**	**9,041.1**	**9,064.5**
Present Value of FCFF	*USD mn*	46,412.1										
Present Value of Terminal Cash Flow	*USD mn*	-										
Total Enterprise Value	*USD mn*	**46,412.1**										
Adjustment for Debt	*USD mn*	29,639.0										
Adjustment for Cash and Marketable Se	*USD mn*	9,375.0										
Investments and Others	*USD mn*	303.0										
Minority Interest	*USD mn*	110.0										
Present Value of Equity	*USD mn*	**26,341.1**										
No. of shares outstanding	*mn*	**1,560.0**										
Value Per Share	*USD*	**16.9**										
Current Price	*USD*	92.5										
Potential Upside/Downside		**-81.7%**										
Calculation of WACC												
Beta	*x*	0.61										

Valuation | Financials | Assumptions | Ratio_Analysis | Graphical_Representation | Technical Analysis | BB

业务假设、跨表引用与文件完整性需要贯穿整个工作流

该任务的核心是处理**跨表依赖**：假设变化影响预测数据，进而传导至后续计算与估值，需确保整个工作簿逻辑一致，并保留原有结构和格式。

Debugging 示例：全面审计与系统化修复

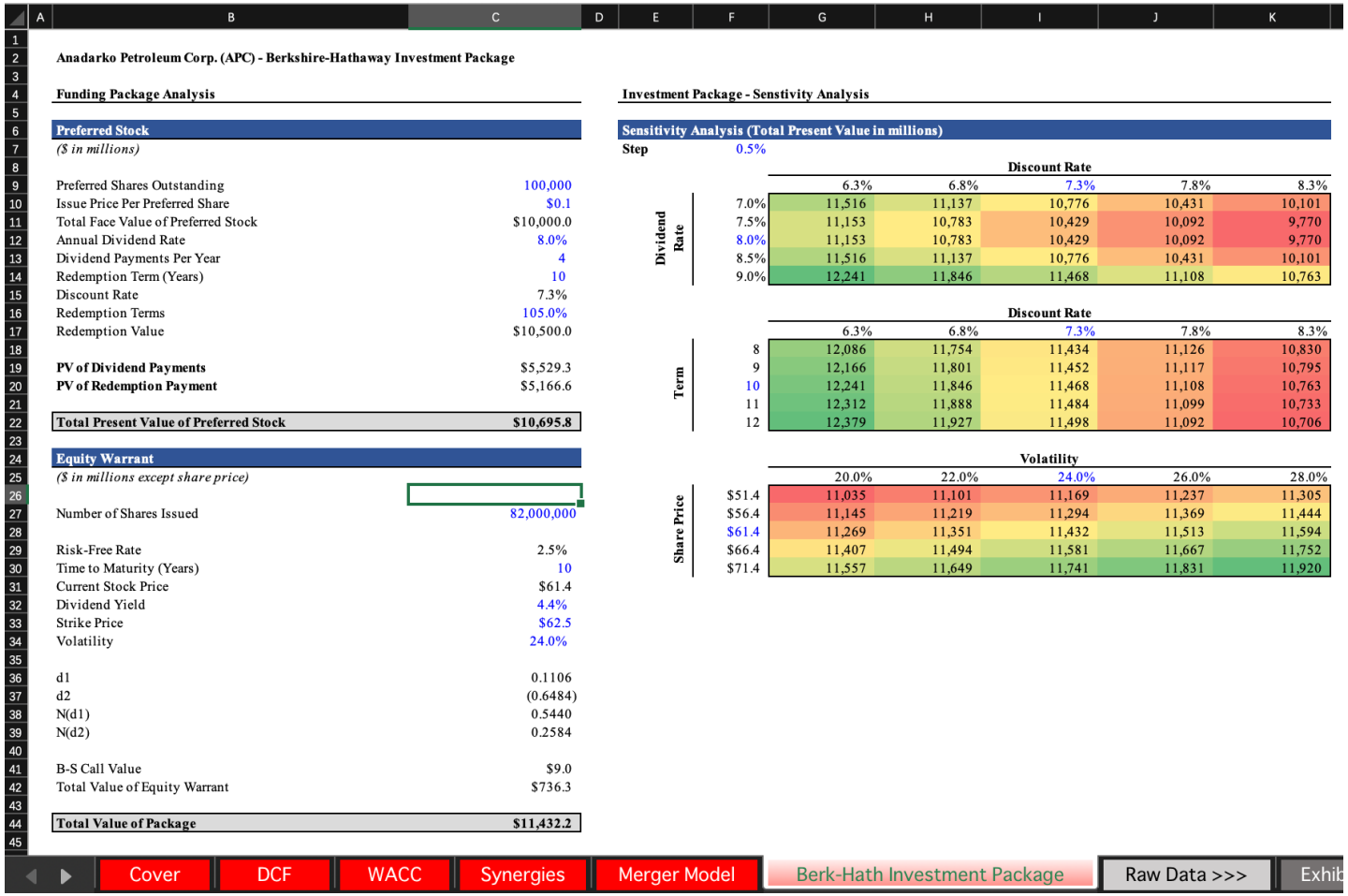
用户给出审计目标，需要智能体主动发现、定位并修复错误

可能的问题

公式逻辑错误或不一致
应使用公式却被硬编码的数值
单元格引用断裂
数据验证与潜在计算错误
格式、命名和最佳实践问题

修复后仍需核验

是否遗漏问题，或引入新的影响



Anadarko Petroleum Corp. (APC) - Berkshire-Hathaway Investment Package

Funding Package Analysis

Preferred Stock	
($ in millions)	
Preferred Shares Outstanding	100,000
Issue Price Per Preferred Share	$0.1
Total Face Value of Preferred Stock	$10,000.0
Annual Dividend Rate	8.0%
Dividend Payments Per Year	4
Redemption Term (Years)	10
Discount Rate	7.3%
Redemption Terms	105.0%
Redemption Value	$10,500.0
PV of Dividend Payments	$5,529.3
PV of Redemption Payment	$5,166.6
Total Present Value of Preferred Stock	**$10,695.8**

Equity Warrant	
($ in millions except share price)	
Number of Shares Issued	82,000,000
Risk-Free Rate	2.5%
Time to Maturity (Years)	10
Current Stock Price	$61.4
Dividend Yield	4.4%
Strike Price	$62.5
Volatility	24.0%
d1	0.1106
d2	(0.6484)
N(d1)	0.5440
N(d2)	0.2584
B-S Call Value	$9.0
Total Value of Equity Warrant	$736.3
Total Value of Package	**$11,432.2**

Investment Package - Senstivity Analysis

Sensitivity Analysis (Total Present Value in millions)

Step 0.5%

Dividend Rate \ Discount Rate	6.3%	6.8%	7.3%	7.8%	8.3%
7.0%	11,516	11,137	10,776	10,431	10,101
7.5%	11,153	10,783	10,429	10,092	9,770
8.0%	11,153	10,783	10,429	10,092	9,770
8.5%	11,516	11,137	10,776	10,431	10,101
9.0%	12,241	11,846	11,468	11,108	10,763

Term \ Discount Rate	6.3%	6.8%	7.3%	7.8%	8.3%
8	12,086	11,754	11,434	11,126	10,830
9	12,166	11,801	11,452	11,117	10,795
10	12,241	11,846	11,468	11,108	10,763
11	12,312	11,888	11,484	11,099	10,733
12	12,379	11,927	11,498	11,092	10,706

Share Price \ Volatility	20.0%	22.0%	24.0%	26.0%	28.0%
$51.4	11,035	11,101	11,169	11,237	11,305
$56.4	11,145	11,219	11,294	11,369	11,444
$61.4	11,269	11,351	11,432	11,513	11,594
$66.4	11,407	11,494	11,581	11,667	11,752
$71.4	11,557	11,649	11,741	11,831	11,920

Cover | DCF | WACC | Synergies | Merger Model | Berk-Hath Investment Package | Raw Data >>> | Exhib

SpreadsheetBench：全球电子表格 Agent 的事实标准评测

50+ 家公司　　联系咨询上榜事宜

模型厂商主动公开评测

V1：OpenAI [1]、Anthropic [2]、Qwen3.7 [3]

V2：Kimi K3 [4]、阶跃星辰 Step 5 [5]

头部办公产品持续参与

Google Sheets [6]、微软 Excel [7]、金山 WPS、字节跳动 [8]

榜单成绩带来行业传播

WPS 曾登顶榜单，获搜狐 [9]、新浪 [10] 等媒体报道

真实业务难题，成为模型能力与办公产品交付能力的共同标尺

BuySideFlow：投研工作流需要时点有效性

能执行、结果能匹配，仍可能用到研究时点尚不可得的信息

真实买方任务与生产库

股票 / 基金 / 固收 / 宏观，覆盖 SQL 与 SQL–Python 完整工作流

任务级时点规则

披露截止日、历史状态与记录版本，约束贯穿检索和计算
一个SQL可以成功执行，输出也可以与参考结果一致，但如果选股时使用了研究日之后披露的财报，或者拿今天的上市状态和行业分类替代历史状态，历史决策仍然无效

评价历史投研智能体，必须同时检查结果与信息使用的时间边界

408 个中文任务

股票 166 / 基金 125
固收 93 / 宏观 24

110 表 / 4,547 列

固定授权金融数据库快照
254 SQL-native
154 SQL–Python

Anonymous authors. (2026). BuySideFlow: A Point-in-Time Benchmark for LLM Agents in Investment Research. ACL submission, anonymous manuscript.

BuySideFlow 样例：同一观察日的三类时间约束

2023-10-31：筛选当日已上市的机械行业公司

Q3 营收同比 > 20%，且 Q3 比 Q2 提高 > 10 个百分点

检查点	任务要求	容易出错的做法	应如何处理
上市状态 （实测错误）	恢复观察日股票池	直接使用当前上市状态 LISTEDSTATE = 1	保留当时已上市、 后来退市的公司
行业分类 （原文示意）	使用当日适用的历史行业	把今天的行业分类 用于过去的筛选	按历史有效区间与 当时可得信息选择分类
披露与版本 （原文示意）	同为 2023-09-30 报告 V1：10-25 披露 V2：11-15 披露	只看报告期末 ENDDATE ≤ 观察日	10-31 截止仅可用 V1 V2 虽覆盖较早报告期， 却尚未公开

输出契约：代码、名称、行业、Q2 / Q3 同比与增幅，增长列保留两位小数

报告期末、披露日、历史有效期与记录版本，必须分别理解和执行

BuySideFlow：结果通过仍不能证明时点有效

9 种配置完成 408 个任务，最高结果准确率仍为 36.8%

Agent	模型	总体	SQL原生	SQL–Python
SWE-agent	GPT-5.5	35.0	38.2	29.9
SWE-agent	DeepSeek-V4-Flash	30.9	33.1	27.3
SWE-agent	Claude-Opus-4.7	29.9	32.7	25.3
SWE-agent	GLM-5.1	19.6	24.4	11.7
SWE-agent	Qwen3.6-Plus	16.2	19.7	10.4
Claude Code	Claude-Opus-4.7	24.0	25.2	22.1
Codex	**GPT-5.5**	**36.8**	**42.1**	**27.9**
Spider-Agent -Lite	GPT-5.5	28.7	31.9	23.4
Spider-Agent -Lite	DeepSeek-V4-Flash	22.3	23.2	20.8

193 / 380 个结果通过样本仍有已确认的 PIT 违规

审计范围：100 道共用任务 × 9 种配置，共 900 次执行

主要时间错误

时间语义混用
当前状态替代历史
历史版本重建错误

核心贡献：同时检查结果正确性与执行过程的时点有效性

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

SELM：技能积累带来的维护难题

外部文本技能不断累积，经验干扰与维护问题也随之出现

技能是注入上下文的文本指南，SELM 优化技能库，目标模型参数保持冻结

经验互相覆盖

为修复当前代码任务而重写共享指南，
可能覆盖数据分析任务需要保留的规则。

相关但反复无效

仅按相似度检索，曾多次失败的技能仍会被选中。
看起来相关，未必有实际使用效果。

删除后又同名生成

低效技能退出后，又从失败轨迹中生成并重新加入。
技能库陷入加入、失败、删除、再加入的循环。

核心问题：如何联合管理技能的调用、修改与退出？

Mengxi Hu, Yuxuan Hu, Jinbo Su, et al. (2026). Selm: Self-Evolving Library Management for Multi-Skill Agent Optimization.
https://arxiv.org/abs/2610.24630

SELM：效用反馈、检索与诊断选择

围绕实际使用效果，决定下次调用什么、优先分析什么

1 效用反馈：执行结果更新历史记录

$$u(s) \leftarrow (1-\alpha)u(s) + \alpha \bar{r}_s,$$

$\bar{r}_s$：本轮使用该技能的任务平均奖励
α = 0.4，兼顾历史记录与近期表现

u(s) 是共同使用时的表现，不能直接视为单技能的因果贡献

2 效用检索：相关性与历史表现共同排序

$$R(s, x_i) = (1-w_u)R_{\text{rel}}(s, x_i) + w_u u(s),$$

R_rel：语义与关键词相关性
同类别 / general 筛选后排序
w_u = 0.6，Top-3，限制注入预算

3 诊断优先级：分析结果出乎预期的轨迹

$$p_i = \left| r_i - \bar{u}(\mathcal{S}_t(x_i)) \right|.$$

可靠技能突然失败，或弱技能意外成功
优先分析差距大者，而非只分析失败

r_i 是实际任务奖励，$\bar{u}$ 是已取技能的平均历史效用

分类修改经选择集验收，低效技能退役，快慢更新整合不同时间尺度的经验

SELM：技能库管理的实验贡献

Claude-Opus-4.6，无 harness，目标模型参数保持冻结

主实验（准确率%，SkillsBench 区分软 / 硬）

基准 / 指标	无技能	Memento-Skills	SELM
SkillsBench（软）	40.4	66.1	**69.8**
SkillsBench（硬）	31.5	48.2	**50.7**
OfficeQA	45.3	59.5	**62.8**
SpreadsheetBench	44.3	54.3	**56.8**
LiveMath	35.2	56.0	**61.3**

组件消融（SpreadsheetBench）

设置	准确率
完整方法	**56.5**
无验证门控	42.5
无效用分数	46.3
检索不使用效用	49.2
无墓地机制	50.4
无 meta 与慢更新	42.7

主表 SpreadsheetBench 为 56.8%，消融完整配置为 56.5%，分别按原表报告

核心贡献：联动管理技能的检索、更新与退出，提升冻结智能体

消融支持效用检索、验证与退役的作用，收益来自整个技能库的维护。

提纲

Contents

01 背景与动机

02 可扩展、高保真的高效数据合成方法

03 稳定高效的强化学习

04 真实业务驱动的复杂任务评测

05 技能自进化

06 总结与展望

总结与展望：把局部改进转化为持续能力增长

已有研究带来的启示

数据价值取决于覆盖能力缺口
从重复出题，转向由当前模型行为决定合成目标

结果正确还需要过程证据支撑
把新证据、环境变化与时间约束纳入训练和验证

交付能力来自模型与外围系统共同作用
工具、验证器与技能库应和模型一起进入评价

部署后自进化的四条研究线

泛化	跨任务、时间与环境留出验证
无标准答案验证	可执行约束、独立证据与影子对照
前瞻演化	上线前预判难题，演化数据与环境
内化与协同	明确内化门槛，对照检验模型 × harness

持续自进化的判据：可泛化、可持续的净收益