面向复杂业务场景的 Deep Research Agent

长程检索架构 · 记忆强化 · Agentic RL 实践

Alibaba Token Hub (ATH) MaaS 应用算法

算法专家 兰天

目录

CONTENTS

Alibaba

真实业务场景应用：需要从"对话"走向"任务完成"

Chatbot 时代

任务：单轮 → 多轮对话
价值：信息提取、知识问答、文本生成
例子：ChatGPT、Gemini

Deep Research Agent 时代

任务：长程规划 + 多源检索 + 工具调用
价值：完成真实业务的任务
例子：Cursor、Claude Code、QoderWork、Qoder

真实业务需要的是一个能带着目标、把信息找全、
把规则判断清楚、把证据链交付出来的 Agent。

真实业务场景为何困难？

Long-horizon Trajectory

决策链长执行不稳定

- 任务和工具调用几百次
- 中间决策高度依赖前序状态
- 决策误判导致的错误难以恢复

Dense Domain Knowledge

领域知识经验密集更新频繁

- 规则体系庞杂、隐含约束多
- 信息散布在大量异构来源中
- 相似场景反复出现、领域知识高度密集、经验难以沉淀

Error Propagation

错误累积、难检查难恢复

- 前序决策错误影响后续流程
- 错误信息隐蔽难以探查
- 验证能力确实导致最终结果不稳定

关键问题：当前最强 Agent 在这类任务上表现如何？差距究竟有多大？

→ 我们需要严格的 Benchmark 来量化这个 gap，并从失败模式中找到突破方向。

Outline：5 项工作 · 1 条主线

相关工作开源：https://github.com/AIDC-AI/Marco-DeepResearch

① 评测分析	② 系统架构改进	③ 记忆自演进	④ Agentic RL	⑤ 业务落地应用
HSCodeComp DeepWideSearch ACL 2026	Table-as-Search ACL 2026	UMEM ICML 2026	Marco DeepResearch 8B-Scale Search Agent 性能优异	业务场景应用 AIBD HSCode自动化 智能审核
真实业务场景 揭示 Agent 缺陷	结构化规划 Harness 稳定长程信息检索	提炼和管理可泛化 记忆支持稳定自演进	驱动训练的 Agentic RL训练	业务应用价值 工作开源共建

评测分析 → 系统架构改进 → 记忆自演进 → 训练强化 → 业务价值

01 来自真实业务的挑战

HSCodeComp + DeepWideSearch：
当前 Agent 在专业层次化规则应用与深宽搜索上的不足

以海关编码（HSCode）为代表的层级规则应用难题

什么是层级规则应用？

问题：很多专业领域的决策**必须严格遵循专家编写的层级规则体系。**

核心特征：

- **层级结构：**从大类→子类→具体条款
- **模糊边界：**自然语言描述规则边界不清晰
- **逻辑交织：**规则间存在排除条款和交叉引用
- **难回溯：**早期分类错误，后续应用全部失败
- **领域知识密集：**高度凝练的领域知识

HSCode: 面向跨境电商海关编码，全球贸易通用标准，决定税率计算正确性

Product title *t*

Green R36S Retro Handheld Video Game Console Linux System 3.5Inch IPS Screen R35s Pro Portable Pocket Video Player 64G 128G 256G

Price *p* 831.55

Currency *u* CNY

Consumer Electronics	Level1 category	**Product categories *c***
Games & Accessories	Level2 category	
Handheld Game Players	Level3 category	

https://www.xxx.com/item/xxxx.html **Product URL *r***

Product Attributes *A*

Attribute	Value	Attribute	Value	Attribute	Value	Attribute	Value
Origin	Mainland China	Bluetooth-compatible	No	Games Type	Retro Games	Screen Type	IPS
Screen Resolution	640x480	Games included	15000+	Display Size	3.5"	WIFI	No
Category	Handheld Game Players	High-concerned Chemical	None	Type of devices	other	Sold in	sell_by_piece
Measurement unit	100000015	Brand Name	Bokidd y	Operating System	Linux	Supporting Language	Brazilian Portuguese, …
RGB lighting joysticks	No	each pack	1	Package	Yes	Battery Capacity	3000 mAh

Product Image *i*

HSCode *y*

9504904000

Chapter
Heading
Sub-Heading
Country-specific

垂类领域通用难题：（1）法律条文→案情判例；（2）医疗编码→医保报销；（3）海关编码→跨境电商。

知识复杂度三层模型

Level 1

开放域数据

网页访问、长文本理解
任务：InfoSeeking Agent
Benchmark：GAIA / BrowseComp / BrowseComp-ZH

Level 2

结构化数据

数据库、知识图谱、领域数据、代码等结构化数据的理解和使用
任务：垂类 Agent 应用，如 Coding Agent，CRM Agent，BI Agent
Benchmark：SWE-Bench、MedBrowseComp、FinSearchComp

Level 3

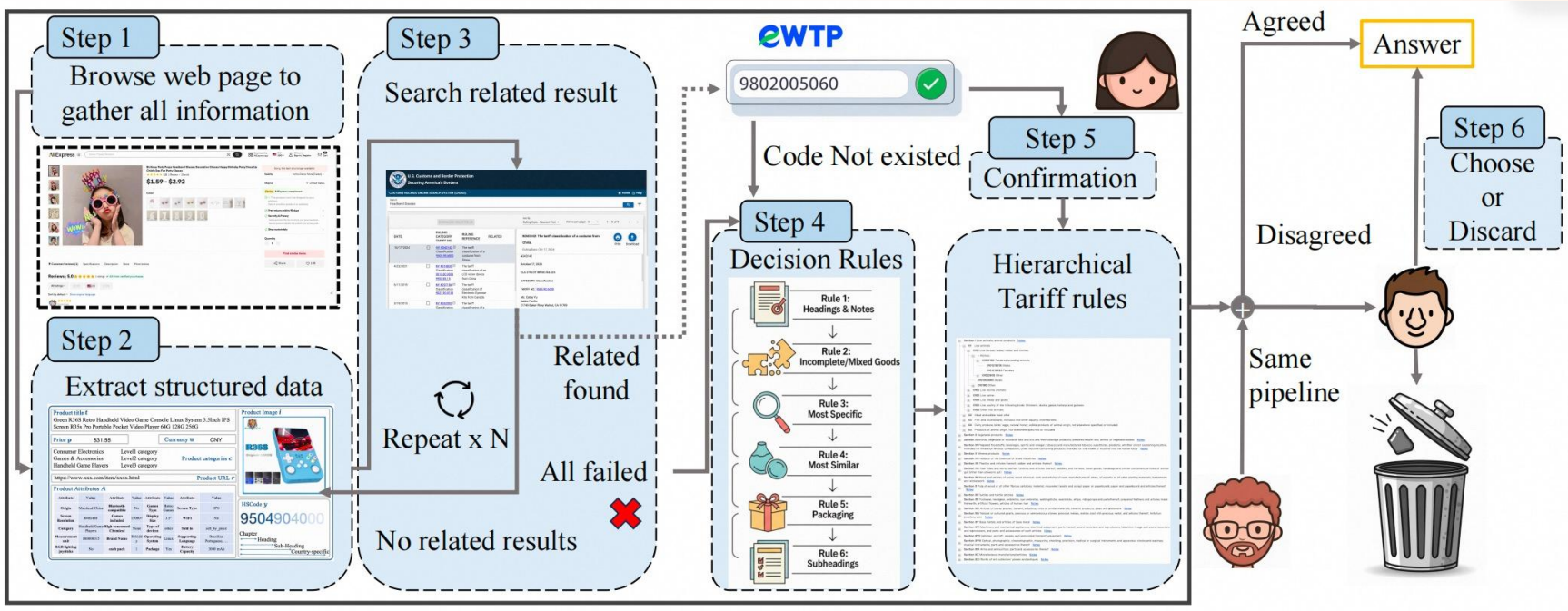
层级规则数据

层次化规则理解和应用
任务：各类高价值垂类任务，如法律、跨境贸易、医疗
Benchmark：HSCodeComp (Ours)

HSCodeComp 数据构建

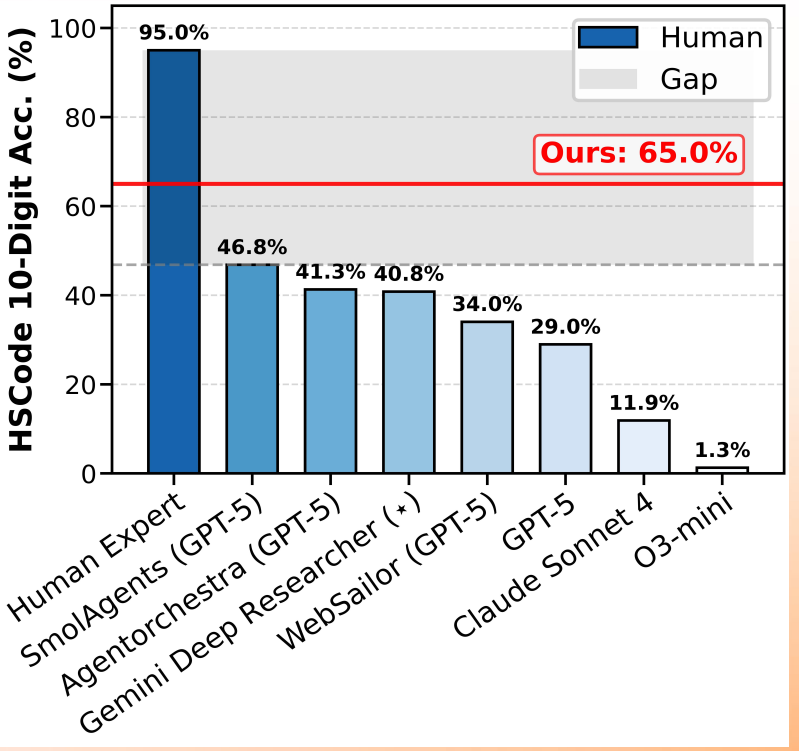
复杂垂类场景Benchmark需要领域专家标注

- **数据规模**：632 个真实电商商品，32 个大类
- **专家标注**：26 位关税专家标注，平均5年+经验
- **规则**：HTSUS 官方税则 + 专家决策规则
- **领域数据**：海关历史裁定记录
- **特点**：真实数据噪声+多模态



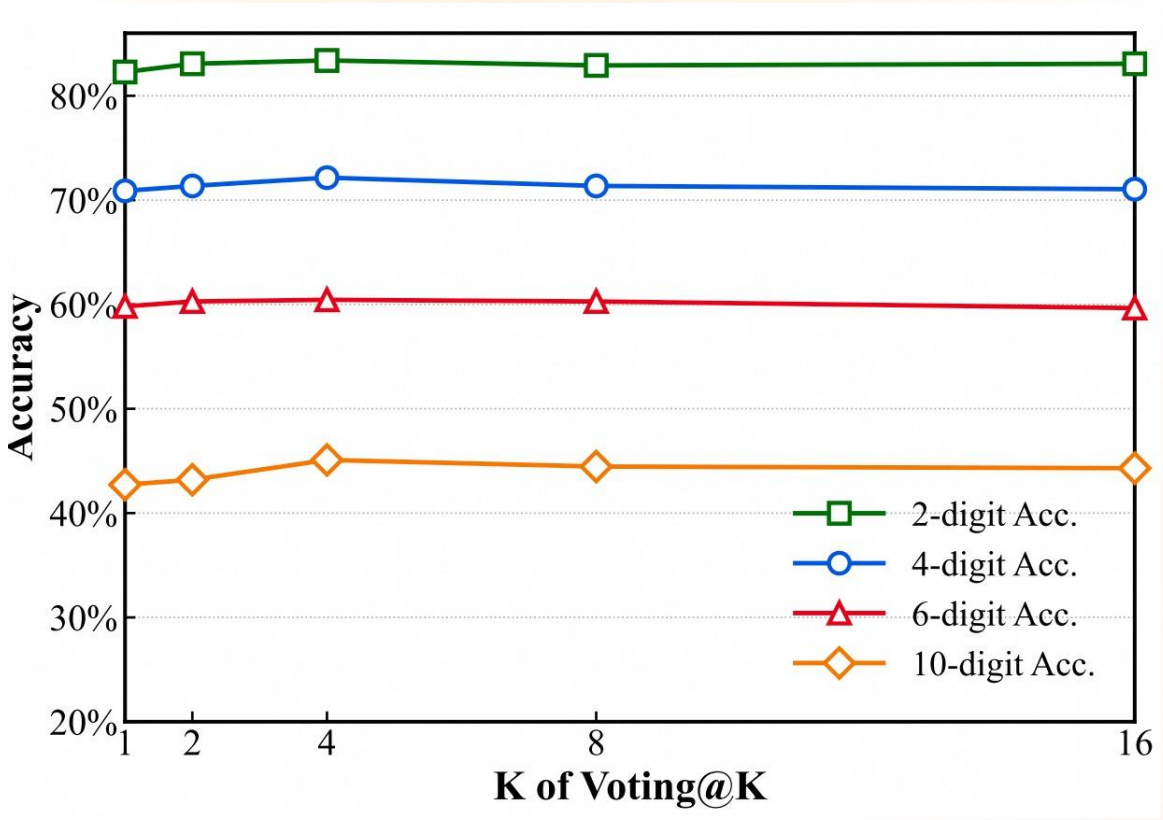
主结果：人类专家 vs. SOTA Agent — 巨大差距

Baselines	Type	HSCode Prediction Accuracy				
		2-digit	4-digit	6-digit	8-digit	10-digit
LLM/VLM-Only						
GPT-5	VLM	82.12	70.89	59.97	41.46	29.27
GPT-5	LLM	82.59	69.78	56.33	40.98	28.96
Gemini-2.5-PRO	VLM	**82.28**	71.04	59.02	40.51	24.21
Gemini-2.5-PRO	LLM	80.54	69.94	58.54	40.35	23.42
GPT-4o	VLM	78.01	64.08	48.10	29.75	18.51
GPT-4o	LLM	75.47	61.55	45.73	30.06	18.35
Claude Sonnet 4	VLM	78.80	64.08	45.25	22.63	11.23
Claude Sonnet 4	LLM	78.80	62.97	44.94	23.58	11.87
Kimi-K2 (Team et al., 2025)	LLM	78.01	62.03	44.15	24.53	12.18
DeepSeek-R1 (Guo et al., 2025)	LLM	77.22	61.71	38.45	16.77	6.65
Qwen-Max (Yang et al., 2025)	LLM	71.52	48.58	24.21	11.23	3.80
O3-mini	LLM	77.22	56.17	24.53	6.65	1.27
Open-source Agent System (GPT-5 Backbone)						
SmolAgents (Roucher et al., 2025)	VLM	82.06	**72.06**	**62.38**	**52.38**	**46.83**
SmolAgents (Roucher et al., 2025)	LLM	**82.28**	70.89	59.81	49.05	42.72
Aworld (Yu et al., 2025c)	LLM	**82.28**	70.41	59.18	48.58	41.30
Agentorchestra (Zhang et al., 2025)	LLM	82.12	70.73	60.44	47.78	41.30
OWL (Hu et al., 2025b)	LLM	72.63	61.87	51.58	41.77	37.34
WebSailor (Li et al., 2025b)	LLM	81.64	70.56	57.27	43.98	35.44
Cognitive Kernel (Fang et al., 2025)	LLM	80.06	69.15	54.59	40.03	26.42



人类专家：95%
最佳 AI Agent：46.8%
针对性优化最佳水平：65%

Test-Time Scaling 策略对层次化规则应用失效



- Majority Voting：K = 1→16 几乎没有改善
- Self-Reflection：降低性能（reasoning drift）
- 核心原因：**在没有可靠的规则和领域知识的辅助、缺乏准确可靠的验证信号情况下，过度的推理和思考会引入幻觉放大错误。**

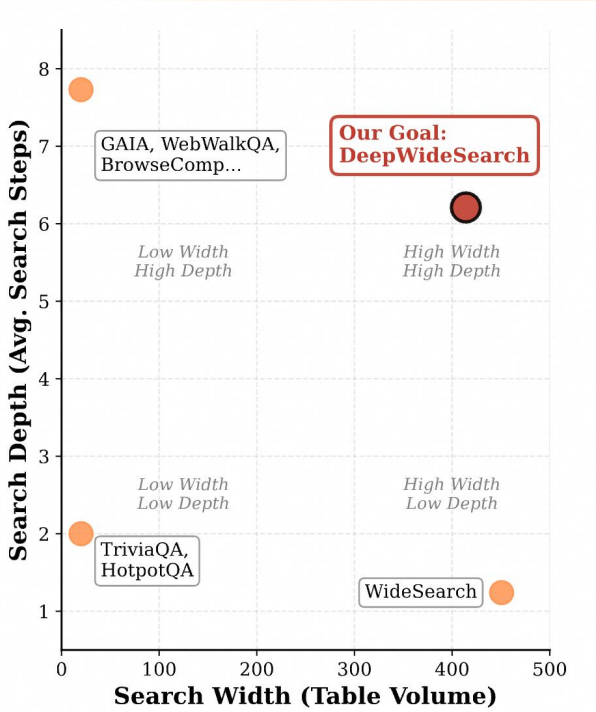
对于层级规则应用："多想几轮"不一定更好。没有可靠领域知识和验证的能力，更多推理会导致错误传播。

HSCodeComp：主要失败模式

过早分类难回溯	在高层类别选择上过早分类进入错误路径、后续难以回溯
错误的自我纠正	缺乏有效垂类知识理解，错误的自我反思反而偏离正确路径
真实噪音干扰	真实场景下，产品描述中的模糊/噪声/错误信息会被干扰 Agent
缺乏领域规则理解	无事实支撑的规则推断、选择无关规则或者遗漏了规则的关联关系

核心挑战是 Agent 对动态领域规则的理解和规则应用过程中的验证纠错

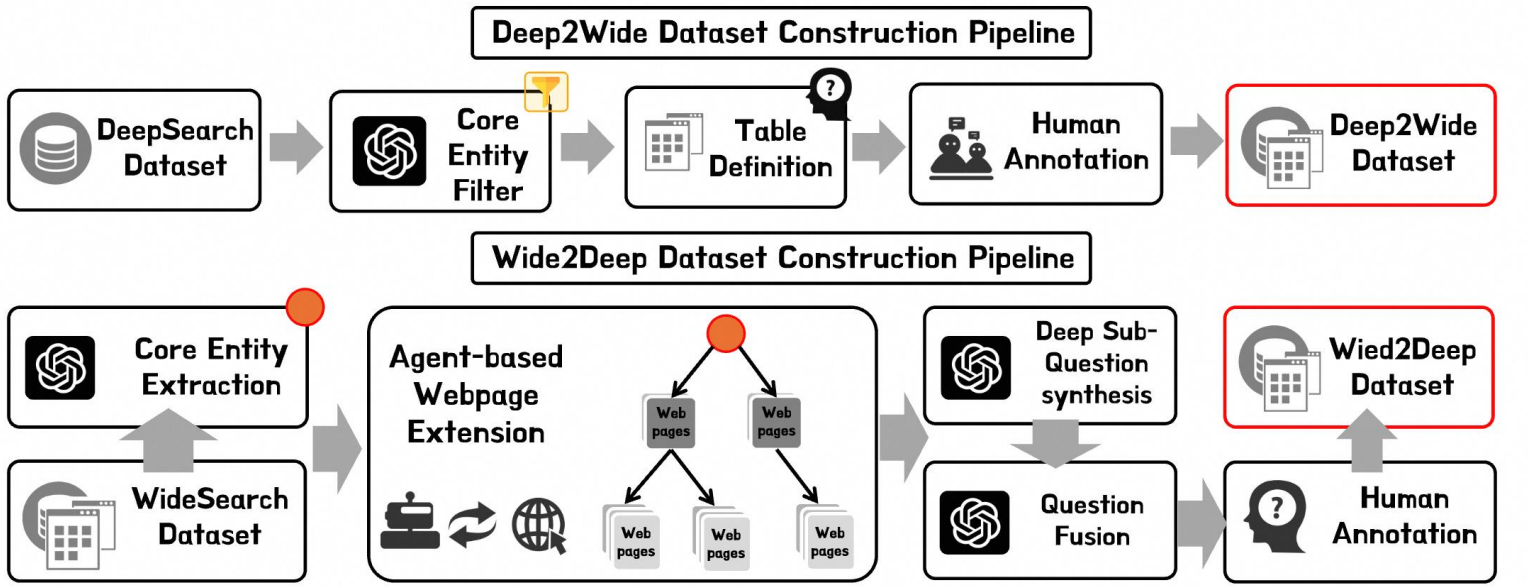
以复杂深宽搜索（DeepWide Search）为代表的长程挑战



Dimension	Deep Search	Wide Search	DeepWideSearch (Ours)
Task	Find specific, hidden answers, facts or entities	Broadly collect target entities and attributes	Broadly collect hidden target entities and attributes that required multi-hop deep web browsing and deep reasoning
Output	A single answer	A structured table of entities and attributes	A structured table of entities and attributes
Challenge	Deep reasoning over multi-step retrieval	Accurately collect a large volume of information about specific entities.	Accurately collect a large volume of information that needed deep reasoning over multi-step retrieval
Example	What is the best-selling EV made by the top company based on MoM sales growth in August 2025?	Collect the Top-3 best-selling new EV cars (price and range) of Geely Inc.	Collect Top-10 EV marker in China by MoM sales growth (Aug 2025) and its Top-3 best-selling new EV cars (price and range).
Human Difficulty	High	Medium	Very High

- InfoSeeking Agent **问题：** （1）Deep Search：强调搜索深度；（2）Wide Search：强调广度信息收集；（3）DeepWide Search：**兼备深度和宽度的组合复杂性**
- **真实挑战：** （1）AIBD场景：找出满足A/B/C/D 所有条件的商家；（2）找到所有今年在 ACL 2026 上发表论文超过5篇的985大学教师或学生作者，并收集其相关信息。

数据构造：Deep2Wide + Wide2Deep + 真实业务数据



Sample A.1: An Example of Our Curated Deep-Wide Search Benchmark

User Query: Please help me identify **30 merchants** that meet all the following criteria:

(1) Target the **Spanish market**;
(2) Sell **Adidas** sneakers;
(3) Offer **competitive pricing**;
(4) Possess mature **B2C operational experience**.

Required Information: For each identified merchant, retrieve the following contact details: *[Phone Number, Cooperation Email, Sales Platform, Official Website, CEO Name, Source URL]*.

- **数据转换：**（1）Deep2Wide：扩展深度搜索范围，将深度搜索转为深宽搜索深度检索 → 扩展为结构化表格信息；
  （2）Wide2Deep：引入多跳验证复杂化，将简单的宽度搜索扩展为深度搜索
- **真实数据标注：**从跨境电商场景的Business Development需求出发构建

主结果：SOTA Agent 成功率极低

Model / System	Success Rate (%)		Row F1 Score (%)		Item F1 Score (%)		Column F1 (%)		CE Acc. (%)	
	Avg@4	Pass@4	Avg@4	Max@4	Avg@4	Max@4	Avg@4	Max@4	Avg@4	Pass@4
Closed-source LLMs										
OpenAI o3-mini	0.0	0.0	3.35	4.55	13.59	16.85	27.36	35.68	61.59	69.55
GPT-5	0.30	1.36	9.61	13.42	21.67	28.21	31.71	41.05	58.41	72.72
Claude Sonnet 4	0.9	0.9	7.31	8.97	19.94	23.38	32.63	40.16	57.95	64.09
Gemini 2.5 Pro	0.9	1.82	15.42	18.96	32.06	37.10	**45.27**	52.86	**73.98**	81.82
Qwen-Max	0.0	0.0	4.16	6.18	14.32	18.48	28.81	36.19	56.02	63.64
GPT-4o	0.0	0.0	4.18	7.01	11.86	16.41	19.66	27.07	54.20	63.64
Open-source LLMs										
DeepSeek-V3	0.23	0.45	6.52	9.99	19.08	24.32	31.26	39.56	60.68	69.09
DeepSeek-R1	0.28	0.45	10.72	14.39	25.01	30.56	38.42	47.77	66.93	75.45
KIMI-K2	0.34	0.91	7.74	11.92	20.44	27.54	31.48	41.83	64.32	73.18
Qwen3-235B-A22B	0.0	0.0	2.94	5.74	12.38	19.53	22.03	34.99	52.39	67.73
Qwen3-235B-A22B-Instruct	0.0	0.0	3.50	5.34	13.28	17.85	24.64	33.03	56.82	64.09
Qwen3-32B	0.0	0.0	2.28	3.67	12.05	16.26	26.37	35.97	54.66	66.36
Open-source Agent Framework with Advanced LLMs										
OWL (Gemini 2.5 Pro)	0.0	0.0	11.11	16.93	28.75	41.70	34.84	50.39	66.14	81.82
OWL (Claude sonnet 4)	0.68	1.36	8.29	14.81	20.44	31.65	30.08	45.50	67.39	81.82
Smolagents (Gemini 2.5 Pro)	0.11	0.45	9.01	15.65	18.53	30.91	27.39	45.09	60.00	79.09
Smolagents (Claude sonnet 4)	0.91	0.91	5.06	8.94	14.49	22.68	21.60	33.83	62.95	74.09
Smolagents (GPT-5)	0.45	0.45	8.18	14.27	20.26	30.66	31.83	44.41	66.48	80.00
WebSailor (Gemini 2.5 Pro)	1.25	2.73	12.51	20.49	25.29	39.11	34.41	52.69	70.57	81.36
WebSailor (Claude Sonnet 4)	**2.39**	**3.64**	**16.88**	**24.26**	**32.90**	**42.35**	42.01	**54.01**	70.91	80.90
WebSailor (GPT-5)	0.34	1.36	10.97	16.17	25.96	35.65	37.18	49.48	74.32	**85.00**

Models / Systems	ReAct	Col-F1	Item-P
Gemini DeepResearch	-	51.2	58.3
Claude-Sonnet-4	SA	39.5	35.2
Claude-Sonnet-4	MA	39.3	44.2

- **任务成功率：**当前最佳的Agent的任务成功率仅为2.39%
- **商业系统不足：**Gemini DeepResearch的列准确率51.2%，单元格准确率58.3%，**无法满足业务需要**

DeepWide Search 四类失败模式

缺少有效反思验证	过早放弃不重试，错误数据难检查难验证
过度依赖内部错误知识	不用网页检索，导致答案过时或错误
缺乏规划浅尝辄止	找到相关网页但没有访问或提取完整信息
上下文窗口爆炸	数百轮长程工具调用导致历史搜索信息挤爆上下文窗口

核心是规划和状态建模：如何在成千上百轮长程搜索中保持干净整洁规划和状态管理。

上下文溢出

真实场景的挑战与我们的解决方案

① 长程任务不稳定

- DeepWideSearch 长程搜索导致上下文迷失
- HSCodeComp 深度推理忽略领域知识

② 经验记忆复用困难

- HSCodeComp的业务知识、规则体系动态更新，模型难适应
- DeepWide Search 的长程信息存储需要高效经验记忆管理

③ 有效验证能力缺失

- HSCodeComp TTS 失效、前序错误难以回溯
- DeepWideSearch 错误信息回填
- 缺乏验证信号、错误传播严重

搜索Agent架构改进

- Table-as-Search：结构化规划 Harness for 信息检索
- 稳定搜索流程支持上千次搜索

自演进记忆管理

- UMEM：提炼管理可泛化记忆
- 提升Agent 记忆库质量，实现稳定的Agent自演进

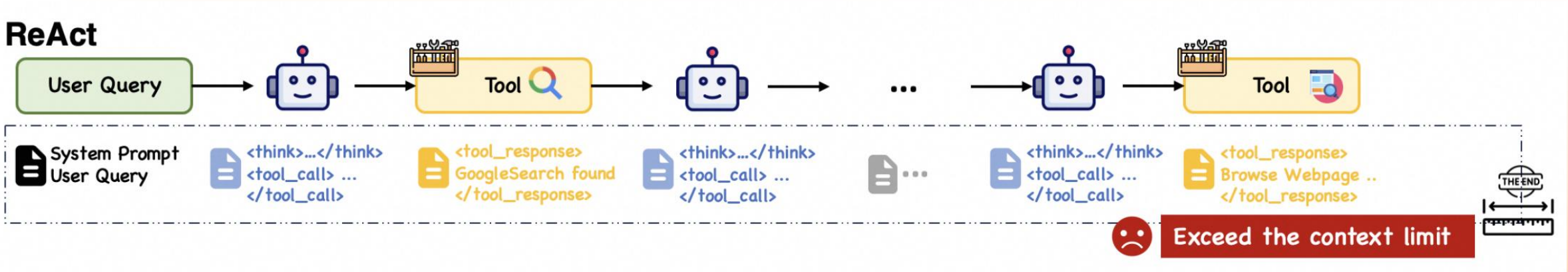
验证驱动的 Agentic RL

- Marco DeepResearch Agent：8B-Scale DR Agent 最佳效果
- 在真实业务场景中得到验证

02 | 长程信息检索 Harness: Table-as-Search

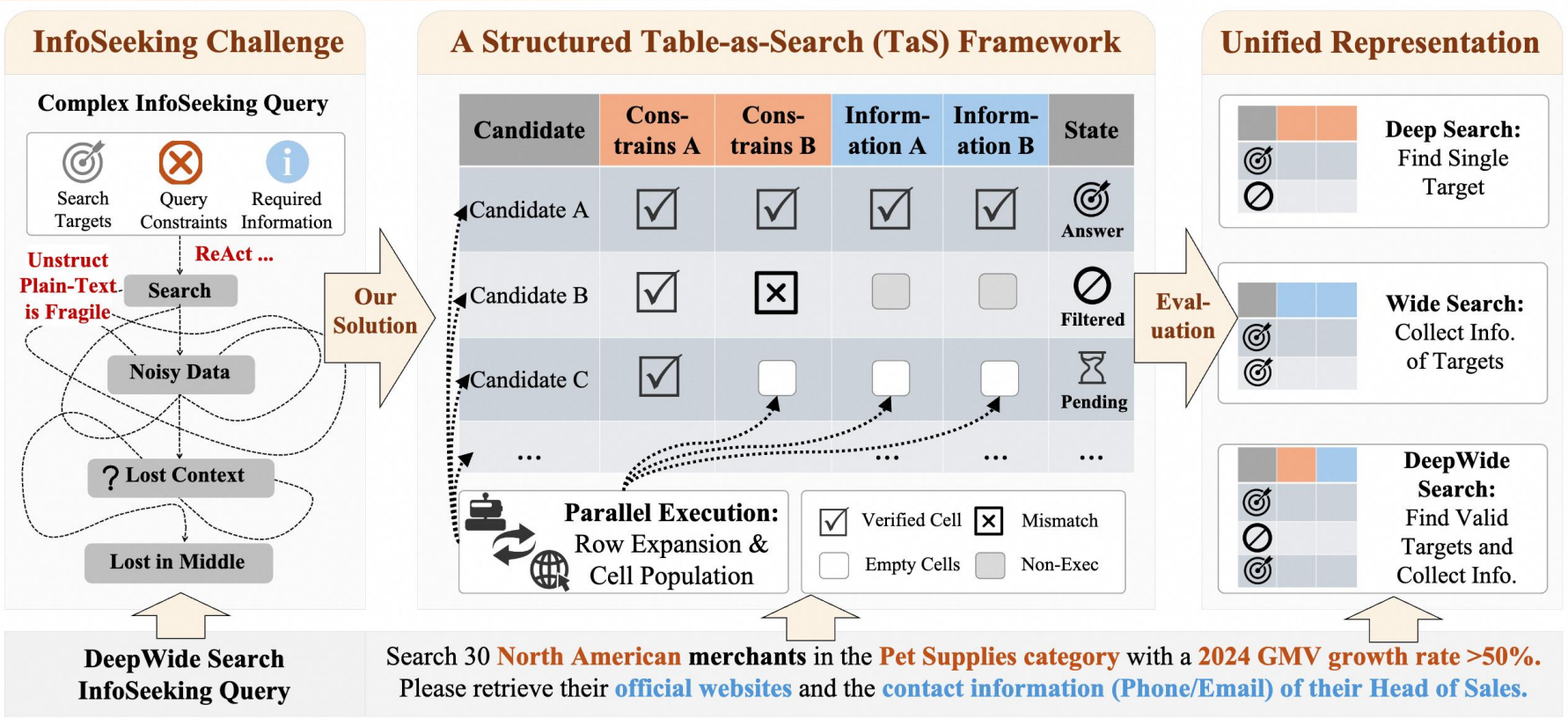
从非结构化规划（ReAct）到结构化规划 Harness

为什么 ReAct 在长程信息搜索任务中非常脆弱



- ReAct Agent：计划 + 搜索结果 + 推理，在无结构文本上下文中
- **信息密度低**：重要信息 lost in the middle
- **状态管理困难**：记规划状态、搜索结果以及推理，认知负担过重
- **扩展困难**：候选增加时性能急剧下降

Table-as-Search 核心：信息搜索 = 表格填充



- Schema <K, C, I>: Keys × Constraints × Information
- **表格即规划**：已填单元格 = 搜索历史；空单元格 = 未来搜索计划
- **表格是高效压缩的搜索状态**：外部数据库承载长程搜索状态
- **表格统一表征三类任务**：Deep Search, Wide Search DeepWide Search

Table-as-Search 长程信息搜索 Harness 架构设计

Algorithm 1: Multi-Agent System of TaS

Input : Query Q, MaxSteps T_{max}, Timeout τ
Output : Final synthesized answer A

```
// Phase 1: Table Initialization
// Define Key Cands./Cons./Info columns
S ← MainAgent.ConstructSchema(Q);
Table ← Initialize(S); State ← Pending
H_T ← {}
// Phase 2: Dynamic Orchestration
while State ≠ Done ∧ ¬Limits(T_max, τ) do
    Plan ← Main.FormulateStrategy(Table, Q)
    if Plan.action == ExpandRows then
        // Case: No enough valid candidates
        {q}_{i=0}^{n} ← MakeQuery(Table.ConsCols)
        foreach q_i in parallel do
            Cands ← SubAgent.DeepSearch(q_i)
            Table.AppendRows(Cands)
    if Plan.action == PopulateCells then
        // Row-Level Parallel Execution
        Rows ← Table.GetIncompleteRows()
        foreach R_i ∈ Rows in parallel do
            q_i ← MakeQuery(R_i, Table.EmptyInfoCols)
            Res_i ← SubAgent.DeepSearch(q_i)
            Table.UpdateRow(R_i, Res_i)
    State ← Main.CheckSaturation(Table)
    Main.Update(H_T)
// Phase 3: Answer Synthesis
return Main.Synthesize(Table, Q)
```

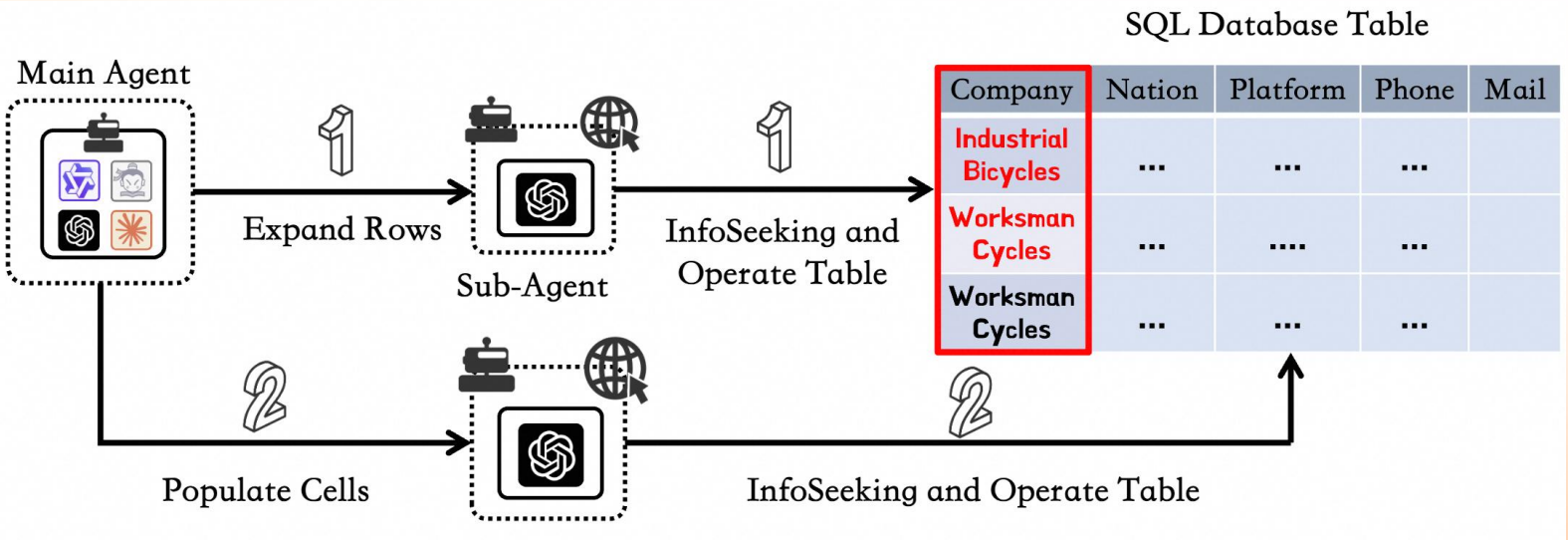


Table Initialization → Dynamic Orchestration → Answer Synthesis

- Phase 1 Table Init：解析 query → 构建 <K,C,I> schema
- Phase 2 Orchestration：Main Agent 动态调度：（1）Row Expansion：发现更多候选实体；（2）Cell Population：对候选逐行验证、补全属性
- Phase 3 Synthesis：从数据库表格检索证据 → 生成结构化答案

主结果：Deep / Wide / DeepWide 三类长程信息搜索任务性能提升

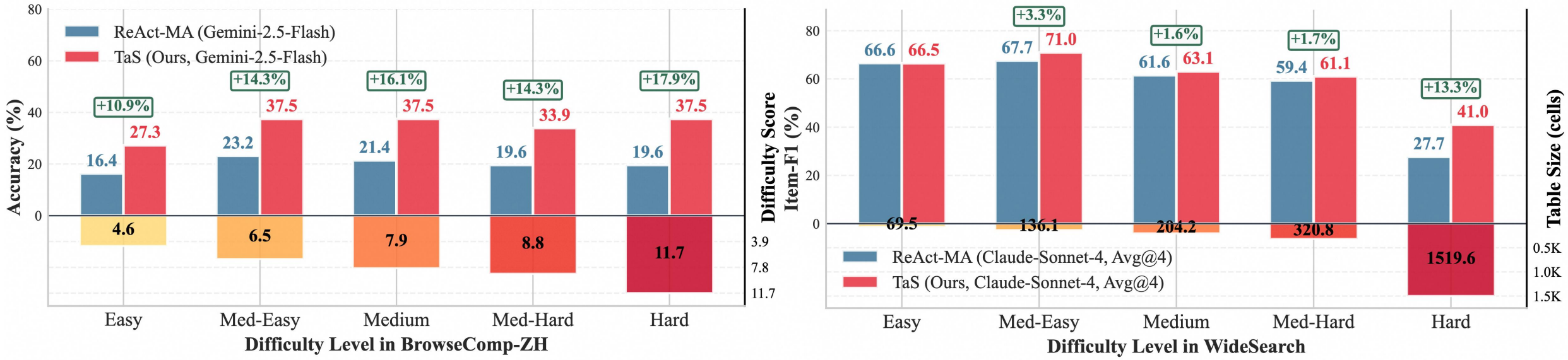
Model / System	Type	GAIA	BC-ZH
Foundation Models with Tools			
OpenAI Deep Research	-	67.4	42.9
GPT-5 High-Think	-	76.4	63.0
Claude-4-Sonnet (Thinking)	SA	68.3	29.1
Gemini-2.5-Pro	SA	60.2	27.8
Training-based Search Agents			
Tongyi DeepResearch (30B)	SA	70.9	46.7
MiroThinker-v1.0-8B	SA	66.4	40.2
MiroThinker-v1.0-30B	SA	73.5	47.8
MiroThinker-v1.0-72B	SA	81.9	55.6
Our proposed TaS Framework			
GPT-5 Medium-Think	SA	66.0	56.5
GPT-5 Medium-Think	MA	71.8	62.9
GPT-5 Medium-Think (Ours)	MA	**77.7**	**63.7**
Qwen3-Max	SA	39.8	23.5
Qwen3-Max	MA	**52.0**	34.3
Qwen3-Max (Ours)	MA	49.0	**35.3**
Gemini-2.5-Flash	SA	16.3	26.6
Gemini-2.5-Flash	MA	38.4	28.4
Gemini-2.5-Flash (Ours)	MA	**52.4**	**34.9**

Model	ReAct Type	SR Acc	Row F1	Item F1	Col F1
Foundation Models with Tools					
Claude-S4 Think	SA	5.0	41.9	66.7	-
Claude-S4 Think	MA	6.5	52.2	73.1	-
Gemini-2.5-Pro	SA	5.0	41.4	63.6	-
Gemini-2.5-Pro	MA	6.5	44.6	66.3	-
OpenAI o3	SA	9.0	44.1	62.3	-
OpenAI o3	MA	9.5	50.5	68.9	-
KIMI-K2	SA	3.5	41.4	65.1	-
KIMI-K2	MA	6.5	49.6	70.7	-
Our proposed TaS Framework					
Gemini-2.5-Flash	SA	**5.0**	41.1	64.8	78.0
Gemini-2.5-Flash	MA	4.5	42.3	61.7	71.4
Gemini-2.5-Flash (Ours)	MA	**5.0**	**45.7**	**67.6**	**82.2**
Claude-S4 NoThink	SA	4.5	38.1	60.9	74.1
Claude-S4 NoThink	MA	4.0	46.8	66.9	78.2
Claude-S4 NoThink (Ours)	MA	**9.1**	**49.0**	**71.0**	**84.4**

Models / Systems	ReAct	Col-F1	Item-P
Gemini DeepResearch	-	51.2	58.3
Claude-Sonnet-4	SA	39.5	35.2
Claude-Sonnet-4	MA	39.3	44.2
Our proposed TaS Framework			
Claude-Sonnet-4 (TaS)	MA	**55.9**	63.5

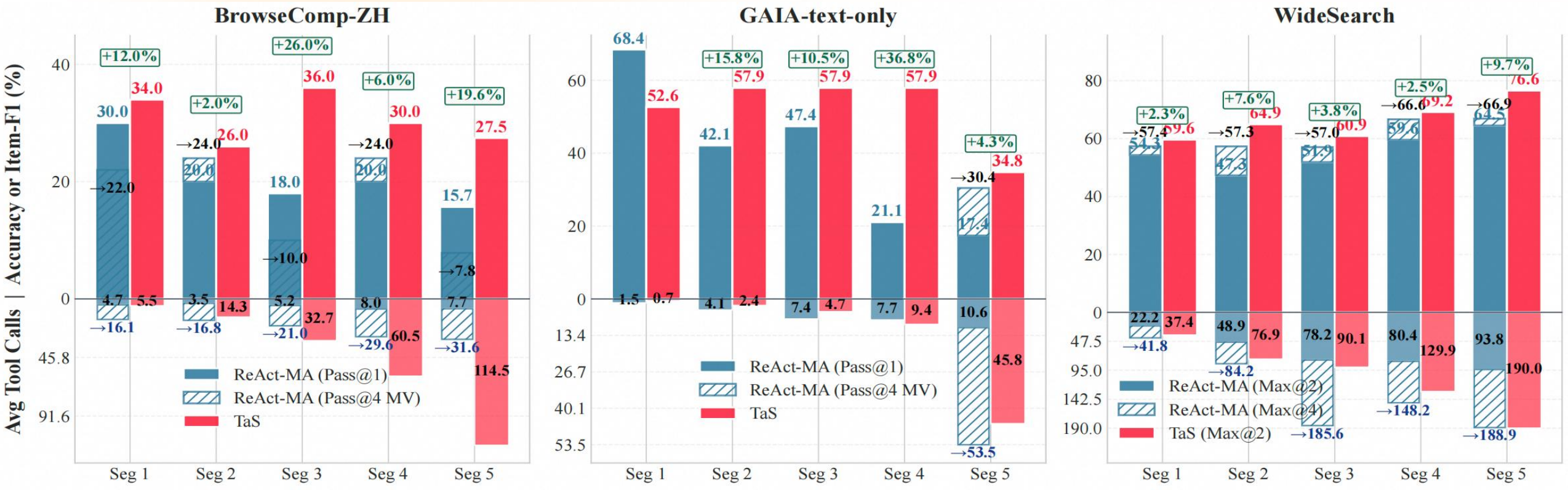
- Deep Search: GAIA +14.0%
- Wide Search: Max@4 9.1%
- DeepWide Search: Col-F1 +4.7% > Gemini DeepResearch

任务复杂度越高，Table-as-Search 优势越明显



- **结构化状态管理防止Long-horizon Agent系统性崩溃**
- BrowseComp-ZH：TaS 优势随着任务复杂持续扩大
- WideSearch：表格超 1500 cells 时，ReAct 退化明显，TaS 下降更慢更稳定

Table-as-Search 搜索效率高



- **搜索效率**：更好结果不完全来自更多工具调用，而来自更精准的搜索计划
- **结果**：更少的工具调用次数有更好的搜索性能

Table-as-Search 扩展性更强

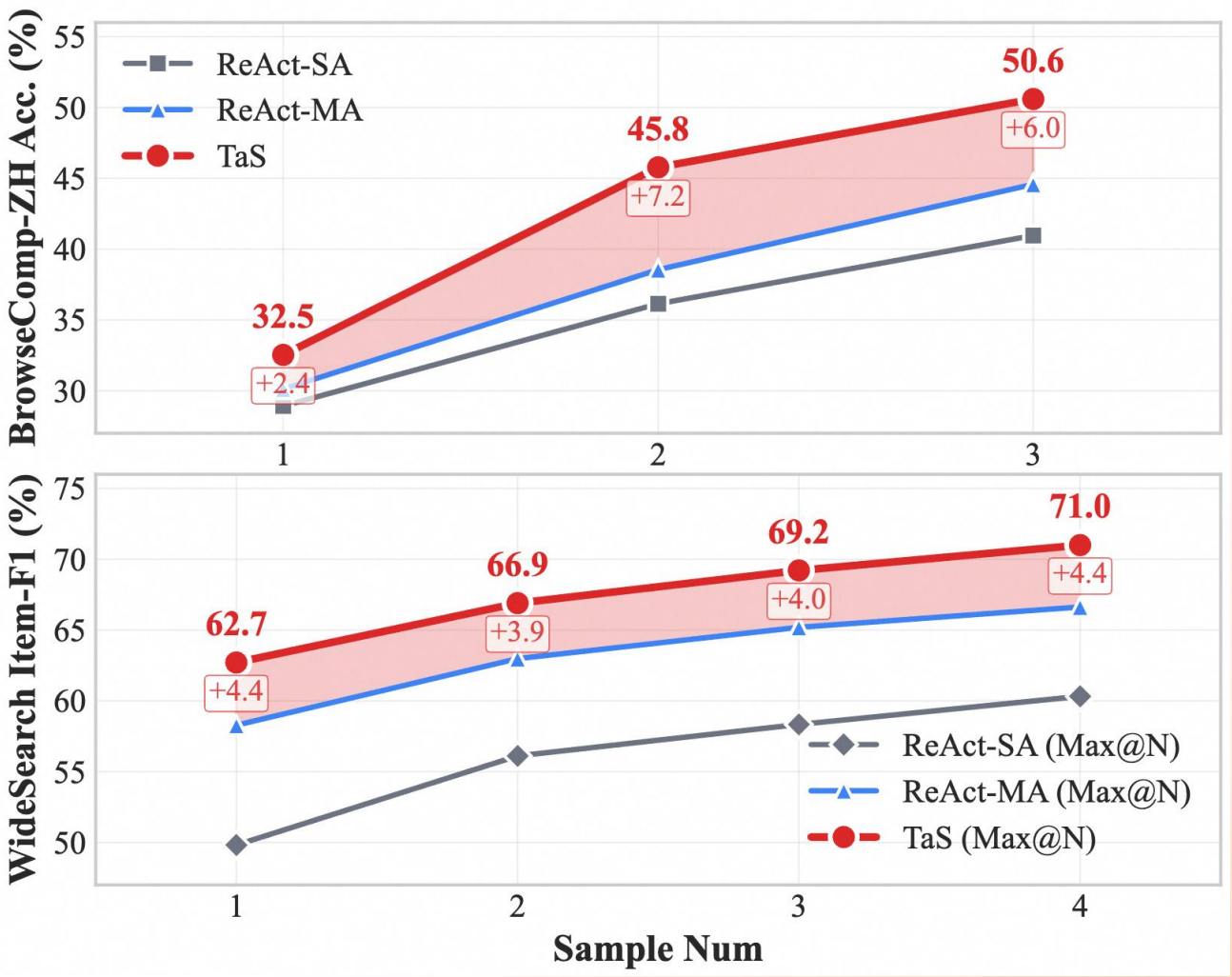


Table-as-Search Test-Time Scaling 更有效

- **扩展性**：深度搜索和宽度搜索稳定优于multi-agent React
- **更稳定**：采样量越大，有时越高

Table-as-Search 更适合工业场景应用

Model Variant	DeepSearch	WideSearch		
	BC-ZH	Row-F1	Item-F1	Col-F1
TaS Framework: `Qwen3-235B-A22B` Sub-Agents. Planners are				
+ Qwen3-Max	**36.5**	**25.5**	**48.0**	58.1
+ Qwen3-235B	29.9	14.6	36.5	50.6
+ Qwen3-30B	7.1	8.6	22.9	33.3
Δ (Qwen3-30B)	**↓29.4%**	↓16.9%	**↓25.1%**	**↓24.8%**
TaS Framework: `Qwen3-Max` Planner. Sub-Agents are				
+ Qwen3-Max	**38.0**	**38.5**	**57.8**	**66.9**
+ Qwen3-235B	36.5	25.5	48.0	58.1
+ Qwen3-30B	27.0	16.9	45.0	63.6
Δ (Qwen3-30B)	↓11.0%	**↓21.6%**	↓12.8%	↓3.3%
TaS Framework: `Gemini-2.5-Flash` Planner. Sub-Agents are				
+ Gemini-2.5-Flash	33.0	**32.7**	52.5	65.8
+ MiroThinker-8B	**40.0**	32.1	**59.0**	**75.9**
Compare with `MiroThinker-v1.0-8B` Standalone Baseline				
Only MiroThinker	32.0	19.8	36.0	47.4
Δ (Only MiroThinker)	↓8.0%	↓12.3%	↓23.0%	↓28.5%

- Table-as-Search 框架中Main-Agent是核心，Main-Agent 影响显著大于Sub-Agent
- 灵活性：替换Sub-Agent为专有搜索Agent（8B-Scale）可以显著降低开销不影响效果

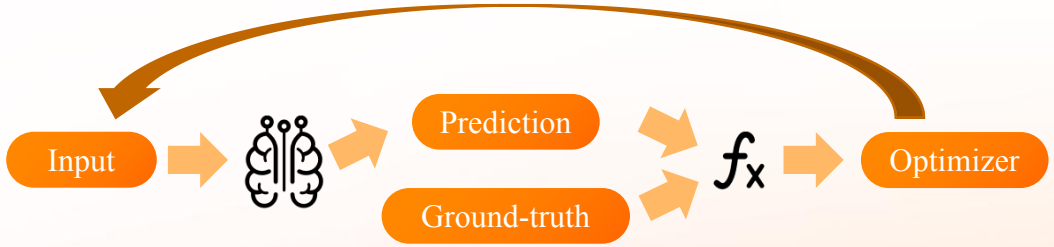
03 | 管理可泛化记忆的自演进：UMEM

统一提取管理可泛化的记忆
让 Agent 从经验中稳定自演进

两类优化方法以及为什么要用记忆自演进

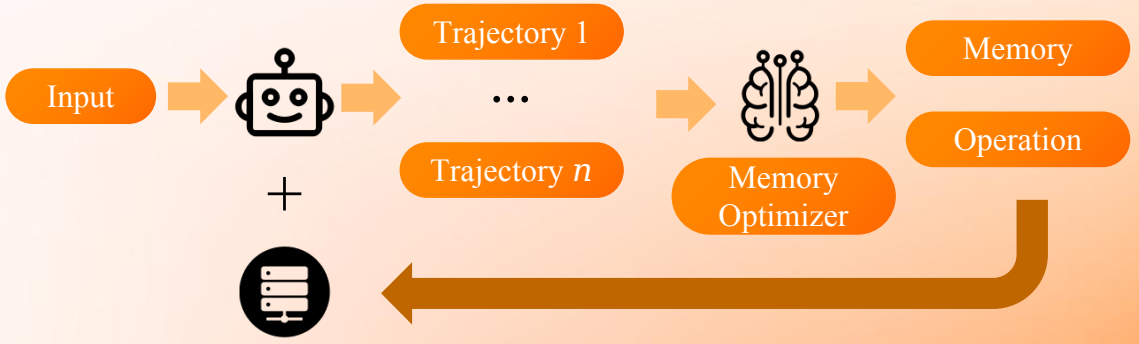
参数化更新模型

- 前向传播 → 计算梯度 → 反向传播
- Loss 计算取决于任务的答案



Self-Evolving Agent Memory 记忆自演进

- Agent 执行获取经验轨迹 → Memory Optimizer Model 提炼经验记忆 + 记忆更新
- 提炼的 Memory 类比梯度



- **复杂业务数据是高频动态更新，固定的模型参数难以适应：** 重新训练 → 成本高、效率低
- **长期记忆 = Agent 的"外挂参数"：** 从真实反馈中沉淀经验 → 未来可复用能力

问题形式化

Self-Evolving Memory: 固定 Executor Agent，优化 Memory Bank

Memory 形式	Memory 检索	轨迹收集（前向推理）	Memory 更新
• Key-Value Memory • Key 是 Query • Value 是提炼的 Memory 文本	• Query → Embedding • 相似度检索 Top-K Key-Value Memory	• Executor Agent 推理，收集轨迹信息 • 获取任务完成的反馈信号	• Mem-Optimizer 从轨迹中提取经验 • 生成 Memory 操作 • 更新Memory Bank

任务：训练一个专有的Mem-Optimizer模型
在流式多任务交互过程中，提炼和管理记忆库，实现 Self-Evolve Agent

现有方法的问题及解决方案

UMEM 联合优化可泛化的记忆提炼和管理流程

现有做法

MemRL, EvolveR, ...

做法：

- 试图在“未来任务”上训练优化
- **但是未来任务的反馈奖励无法传递给过去的训练过程**

问题：

- 记忆管理策略和记忆提炼环节**脱节**
- **未能成功显式**建模记忆的**可泛化性**

UMEM

提取管理可泛化记忆

做法：

- **联合建模记忆提炼和管理：** 端到端生成任务

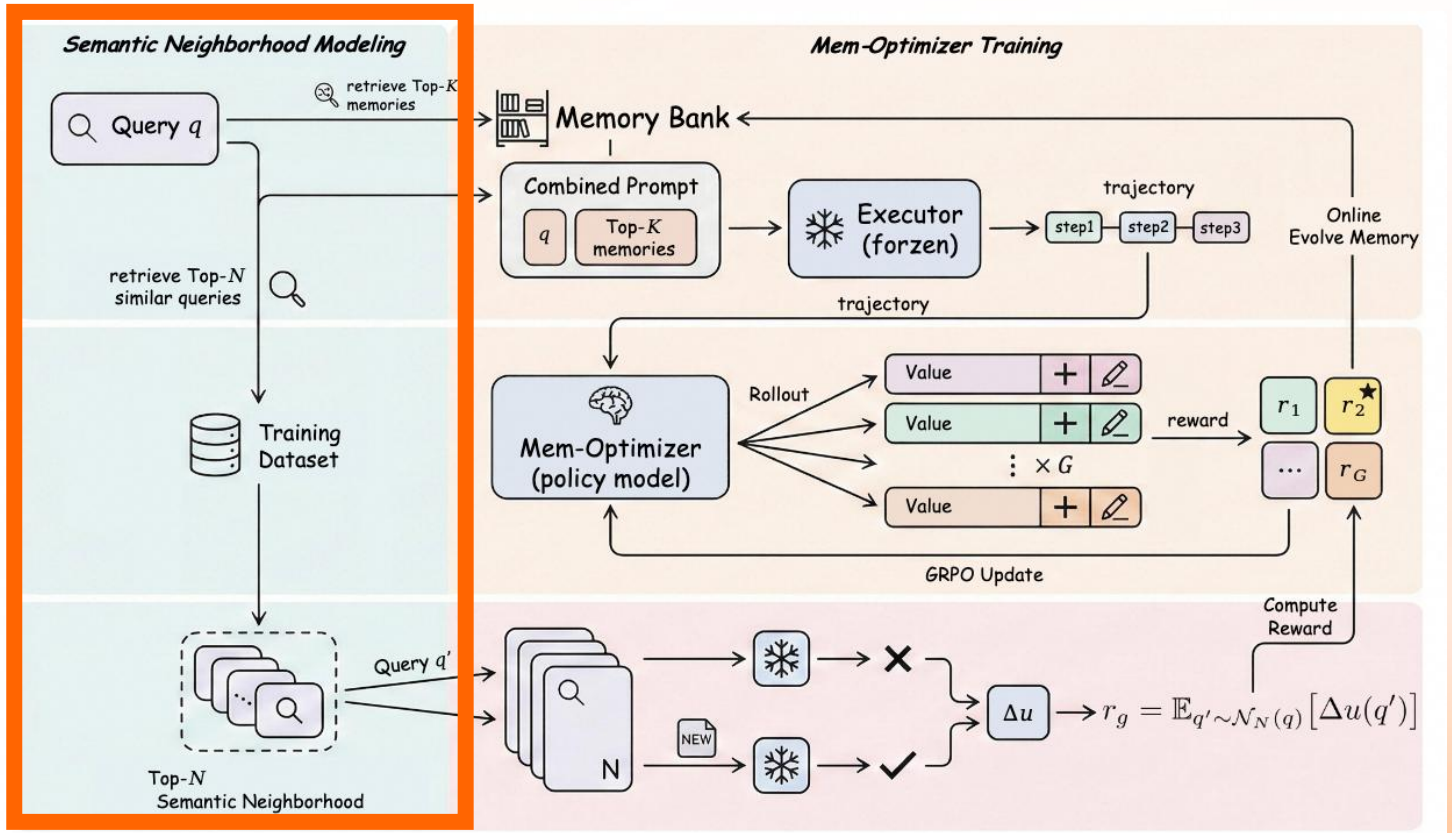
<experience><value>...</value>

<operation>...</operation></experience>

- **直接优化** Memory **可泛化性**
- Mem-Optimizer 与 Bank **共同演化**

UMEM框架：联合提炼和管理可泛化的记忆

直接优化“未来任务”太难；UMEM 用相似任务近似“未来任务”

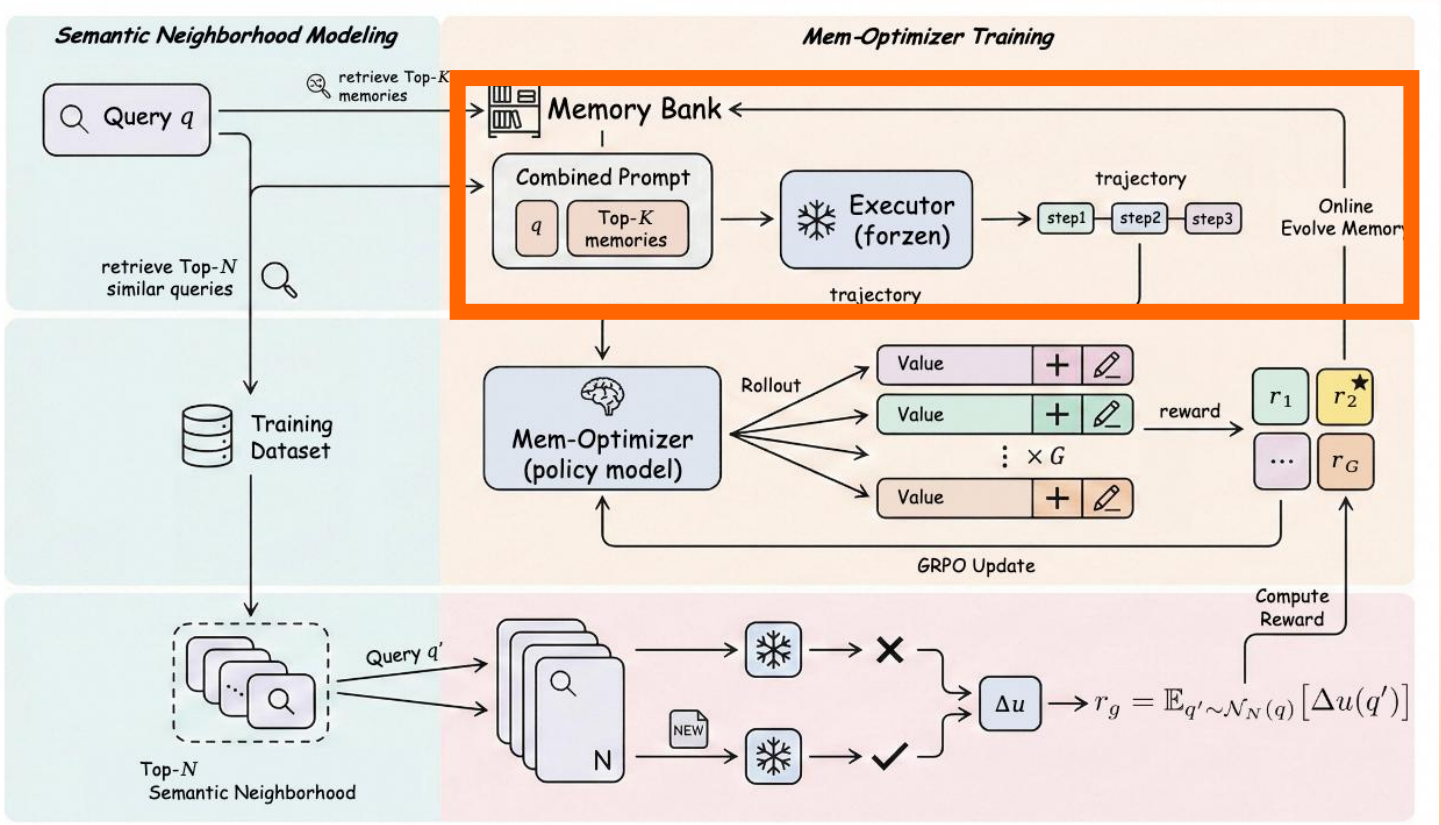


Semantic Neighborhood Modeling

- Query q 映射为语义向量
- 检索 Top-K 相似任务及其 Memory
- 训练过程中持续累积 Memory Bank

UMEM框架：联合提炼和管理可泛化的记忆

直接优化“未来任务”太难；用相似任务近似“未来任务”

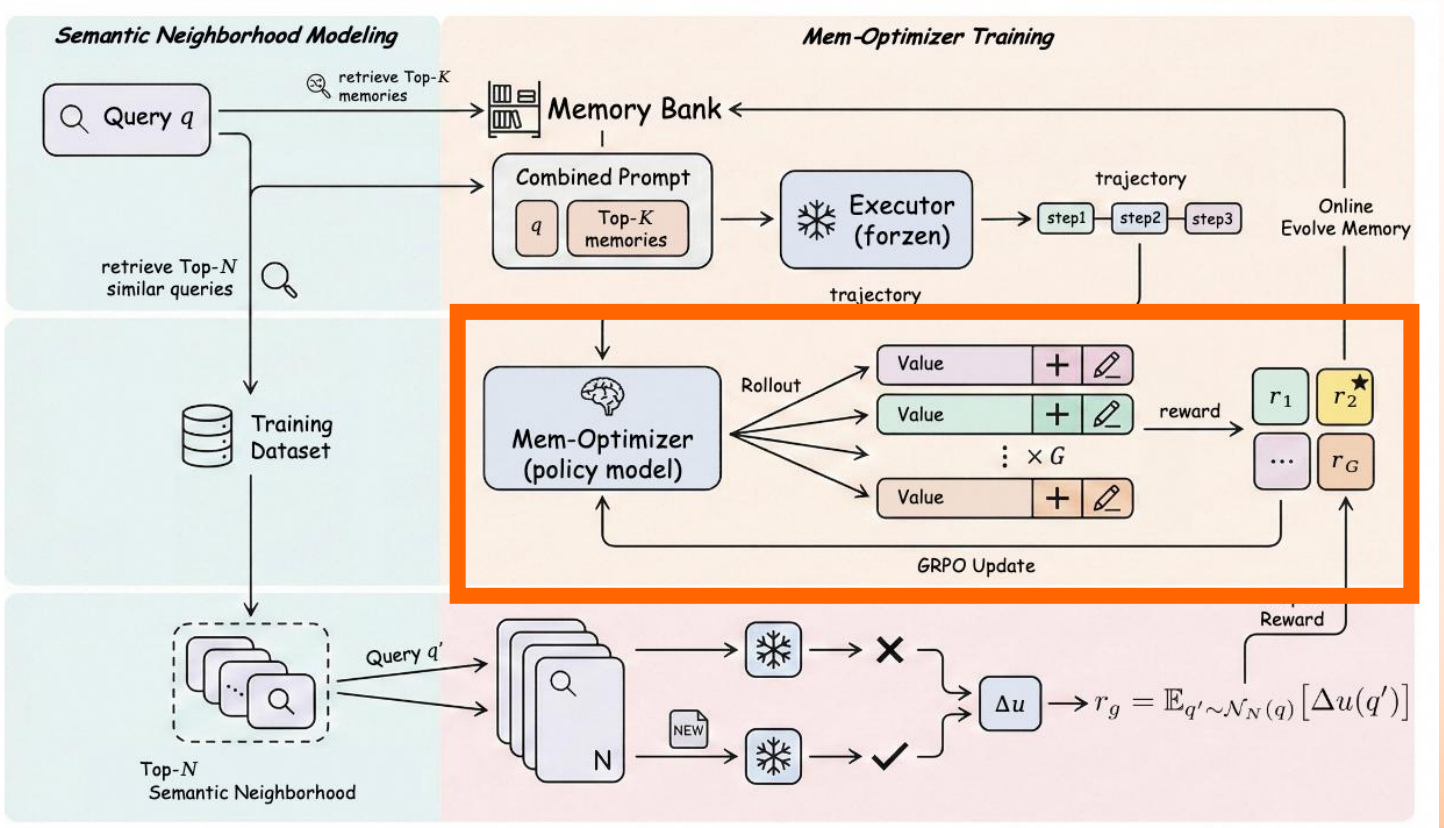


记忆增强的执行和轨迹收集

- 拼接 Query q 和 Top-K 检索记忆
- Memory-Augment Execution: Agent 推理得到若干轨迹数据
- Executor Agent 不参与训练

UMEM框架：联合提炼和管理可泛化的记忆

直接优化“未来任务”太难；用相似任务近似“未来任务”



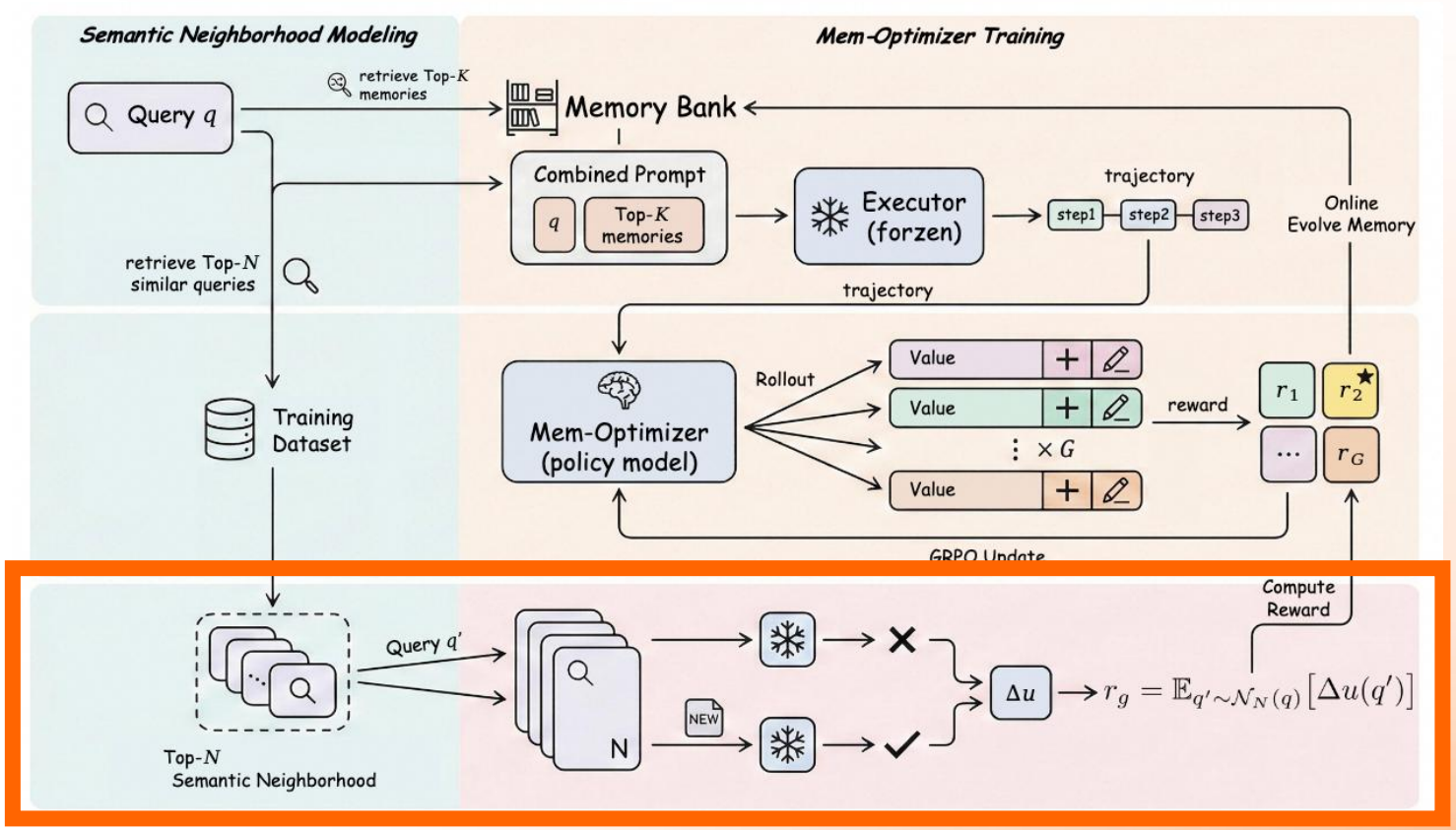
Mem-Optimizer 策略采样

- Mem-Optimizer (UMEM) 从轨迹中提炼记忆及对应操作

$$\{a_q^{(g)}\}_{g=1}^{G} \sim \pi_\phi(\cdot \mid q, \tau_q, \mathcal{B}_t^{topk})$$

UMEM框架：联合提炼和管理可泛化的记忆

直接优化“未来任务”太难；用相似任务近似“未来任务”



面向可泛化记忆的奖励设计

- 相似任务验证：使用/不使用提炼记忆计算重新推理相似任务
- 计算在相似任务上的收益 Reward
- 成功率收益：

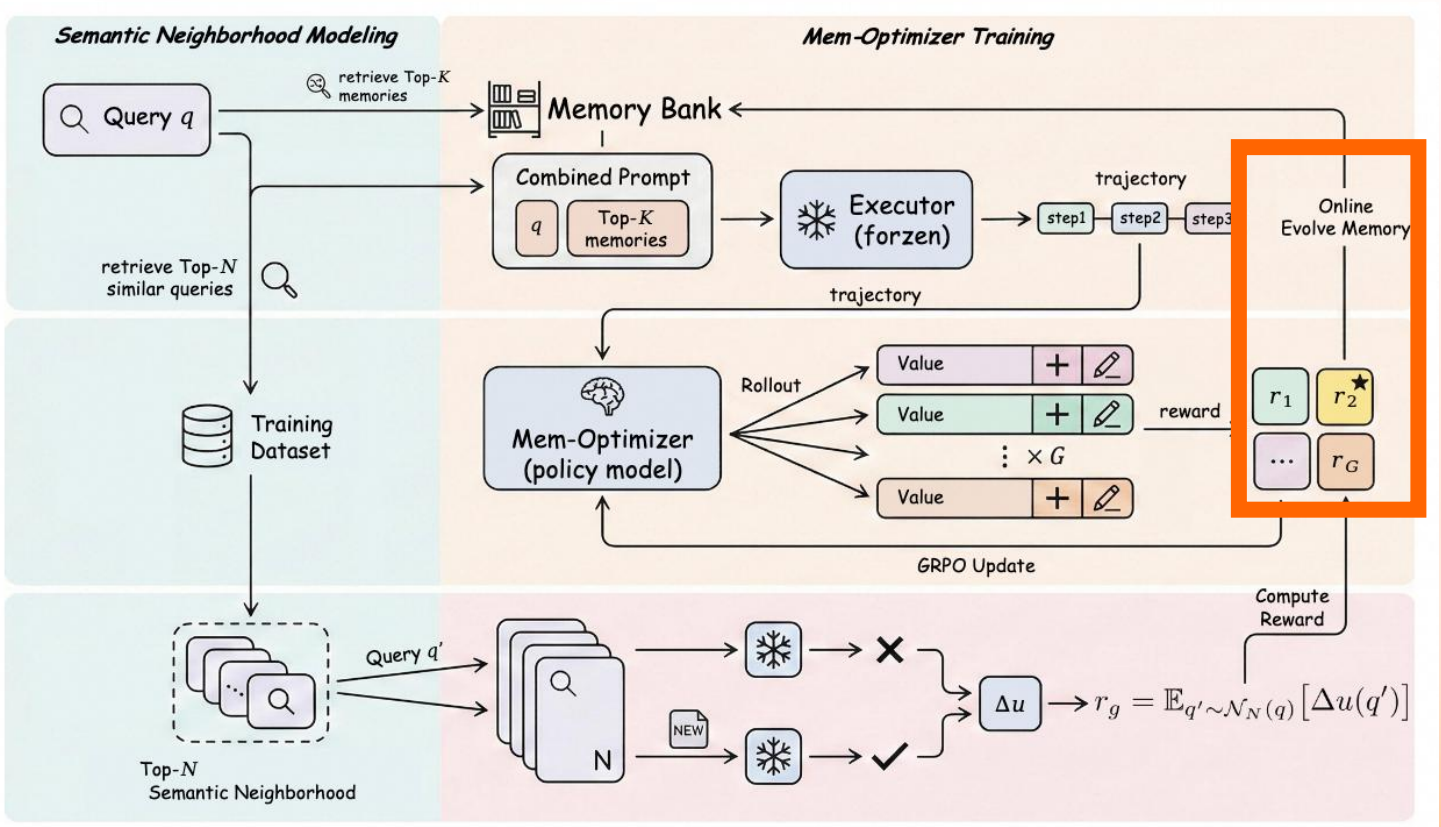
$$\mathcal{G}_{\text{succ}}(q') = c(\tau_{q'}^{\text{mem}}) - c(\tau_{q'}^{\text{ref}})$$

- 执行效率：

$$\mathcal{R}_{\text{eff}}(q') = (c_{\text{mem}} \cdot c_{\text{ref}}) \cdot \left(1 - \frac{\ell_{\text{mem}}^{(g)}}{\ell_{\text{ref}}}\right)$$

UMEM框架：联合提炼和管理可泛化的记忆

直接优化“未来任务”太难；用相似任务近似“未来任务”



持续记忆演化

- 选择当前 GRPO 组内最佳提炼记忆和策略，更新 memory bank
- 拒绝静态memory上训练，随着训练过程持续演化 memory bank
- 让模型学会适应动态变化的 Memory Bank

实验分析：主实验

Models	Parameters	AIME	GPQA	HLE	HotpotQA	ALFWorld		Average
						SR	PR	
Frozen Executor: Qwen3-8B-Thinking								
No Memory	-	51.67	52.53	7.51	62.00	41.04	68.91	47.28
Self-RAG	8B	30.00	43.94	8.56	25.00	30.60	54.23	32.06
ReMem	8B	**61.67**	53.54	6.42	15.00	46.27	66.17	41.51
Memp	8B	46.67	49.49	**11.22**	62.00	44.78	69.78	47.32
Llama-3.2-1B-Instruct	1B	51.67	52.02	6.42	59.00	47.01	74.13	48.38
UMEM-Llama-3.2-1B (Ours)	1B	60.00$_{\uparrow 8.3}$	**54.04**$_{\uparrow 2.0}$	6.95$_{\uparrow 0.5}$	61.00$_{\uparrow 2}$	44.78$_{\downarrow 2.2}$	72.14$_{\downarrow 2.0}$	49.82$_{\uparrow 1.4}$
Qwen3-4B-Instruct	4B	55.00	53.54	6.95	60.00	40.30	65.92	46.95
UMEM-Qwen3-4B (Ours)	4B	58.33$_{\uparrow 3.3}$	52.02$_{\downarrow 1.5}$	8.02$_{\uparrow 1.1}$	**63.00**$_{\uparrow 3}$	**50.75**$_{\uparrow 10.5}$	**73.13**$_{\uparrow 7.2}$	**50.88**$_{\uparrow 3.9}$
Frozen Executor: GPT-5.1								
No Memory	-	40.00	57.57	6.95	39.00	61.94	66.67	45.36
Self-RAG	API	50.00	57.58	7.49	42.00	70.90	83.71	51.95
ReMem	API	30.00	62.63	8.56	43.00	73.13	79.60	49.49
Memp	API	45.00	62.12	**10.16**	52.00	77.61	81.34	54.71
Llama-3.2-1B-Instruct	1B	43.33	61.11	7.49	51.00	61.94	73.63	49.75
UMEM-Llama-3.2-1B (Ours)	1B	45.00$_{\uparrow 1.7}$	62.63$_{\uparrow 1.5}$	8.56$_{\uparrow 1.1}$	**55.00**$_{\uparrow 4}$	64.18$_{\uparrow 2.2}$	75.37$_{\uparrow 1.7}$	51.79$_{\uparrow 2.0}$
Qwen3-4B-Instruct	4B	46.67	62.63	8.02	52.00	70.90	78.86	53.18
UMEM-Qwen3-4B (Ours)	4B	**51.67**$_{\uparrow 5.0}$	**65.15**$_{\uparrow 2.5}$	8.56$_{\uparrow 0.5}$	54.00$_{\uparrow 2.0}$	**82.84**$_{\uparrow 11.9}$	**84.20**$_{\uparrow 5.3}$	**57.74**$_{\uparrow 4.6}$
Frozen Executor: Gemini-2.5-Flash								
No Memory	-	53.33	73.23	10.16	30.00	55.22	72.89	49.14
Self-RAG	API	56.67	71.72	10.16	42.00	59.70	74.50	52.46
ReMem	API	56.67	70.20	10.70	36.00	56.72	75.62	50.99
Memp	API	53.33	74.75	7.49	41.00	60.45	76.74	52.29
Llama-3.2-1B-Instruct	1B	51.67	73.23	10.16	42.00	53.73	71.27	50.34
UMEM-Llama-3.2-1B (Ours)	1B	58.33$_{\uparrow 6.7}$	71.72$_{\downarrow 1.5}$	**13.37**$_{\uparrow 3.2}$	42.00	58.96$_{\uparrow 5.2}$	77.24$_{\uparrow 6.0}$	53.60$_{\uparrow 3.3}$
Qwen3-4B-Instruct	4B	56.67	72.22	9.63	42.00	52.24	72.64	50.90
UMEM-Qwen3-4B (Ours)	4B	**60.00**$_{\uparrow 3.3}$	**76.26**$_{\uparrow 4.0}$	11.76$_{\uparrow 2.1}$	**45.00**$_{\uparrow 3.0}$	**61.19**$_{\uparrow 9.0}$	**78.61**$_{\uparrow 6.0}$	**55.47**$_{\uparrow 4.6}$

Qwen3-4B 版本上优化的 UMEM 在三个 executor **上取得最优表现**

UMEM-Qwen3-4B 提炼记忆优于 ReMem和 Memp +闭源模型 (GPT-5.1 & Gemini-2.5-Flash)

平均提升 3.03% 和 3.18%

实验分析：更多的baseline对比

Method	GPT-5.1	Gemini-2.5-Flash
No Memory	45.36	49.14
ReasoningBank	53.76 ±0.82	54.39 ±0.59
MemRL	45.91 ±0.59	51.67 ±0.87
EvolveR	55.34 ±0.58	52.42 ±0.47
Memp	54.24 ±0.73	51.60 ±0.58
Qwen3-4B-Instruct	53.67 ±0.49	51.00 ±0.91
UMEM-Qwen3-4B (Ours)	**58.10 ±0.72**	**55.46 ±0.82**

在5个benchmark上的平均结果显示：**UMEM-Qwen3-4B 持续稳定优于 SOTA Memory 系统**

实验分析：UMEM 可扩展性分析

Model	Param	AIME	GPQA	HLE	Hotpot	ALF SR	ALF PR	Avg
FROZEN EXECUTOR: QWEN3-8B-THINKING								
No Memory	-	51.67	52.53	7.51	62.00	41.04	68.91	47.28
Llama-3.2-1B	1B	51.67	52.02	6.42	59.00	47.01	74.13	48.38
UMEM-1B (Ours)	**1B**	**60.00** ↑8.3	**54.04** ↑2.0	**6.95**	**61.00**	**44.78**	**72.14**	**49.82**
Qwen3-4B	4B	55.00	53.54	6.95	60.00	40.30	65.92	46.95
UMEM-4B (Ours)	**4B**	**58.33** ↑3.3	**52.02**	**8.02** ↑1.1	**63.00** ↑3.0	**50.75** ↑10.5	**73.13** ↑7.2	**50.88**
Qwen3-14B	14B	57.22	53.71	7.84	61.67	47.51	70.69	49.77
UMEM-14B (Ours)	**14B**	**60.56** ↑3.3	**55.05** ↑1.3	**8.74** ↑0.9	**63.67** ↑2.0	**52.24** ↑4.7	**74.55** ↑3.9	**52.47**

- UMEM-1B 和 4B 模型在部分任务上出现**性能下降**
- 随着 UMEM 尺寸增加逐渐改善；UMEM-14B 模型提炼记忆的能力**持续增强，始终优于基线**

实验分析：跨任务泛化分析

Method	single-bench avg	cross-bench avg
UMEM-4B	57.74	58.33
No-Train	53.18	52.92
No-Memory	45.36	45.36

CTU 和 ERR 对比分析

Method	Cross-Task Utility (CTU)	Error Recovery Rate (ERR)
UMEM-4B	41.47%	23.17%
No-Train	39.29%	17.80%

- 跨任务场景下性能进一步提升，GPT-5.1 Executor 下从 57.74% → 58.33%； Baseline 表现出性能损失
- Cross-Task Utility (CTU) 和 Error Recovery Rate (ERR) 证明 UMEM记忆跨任务可用性更高

04

验证驱动的Agentic RL
Marco DeepResearch

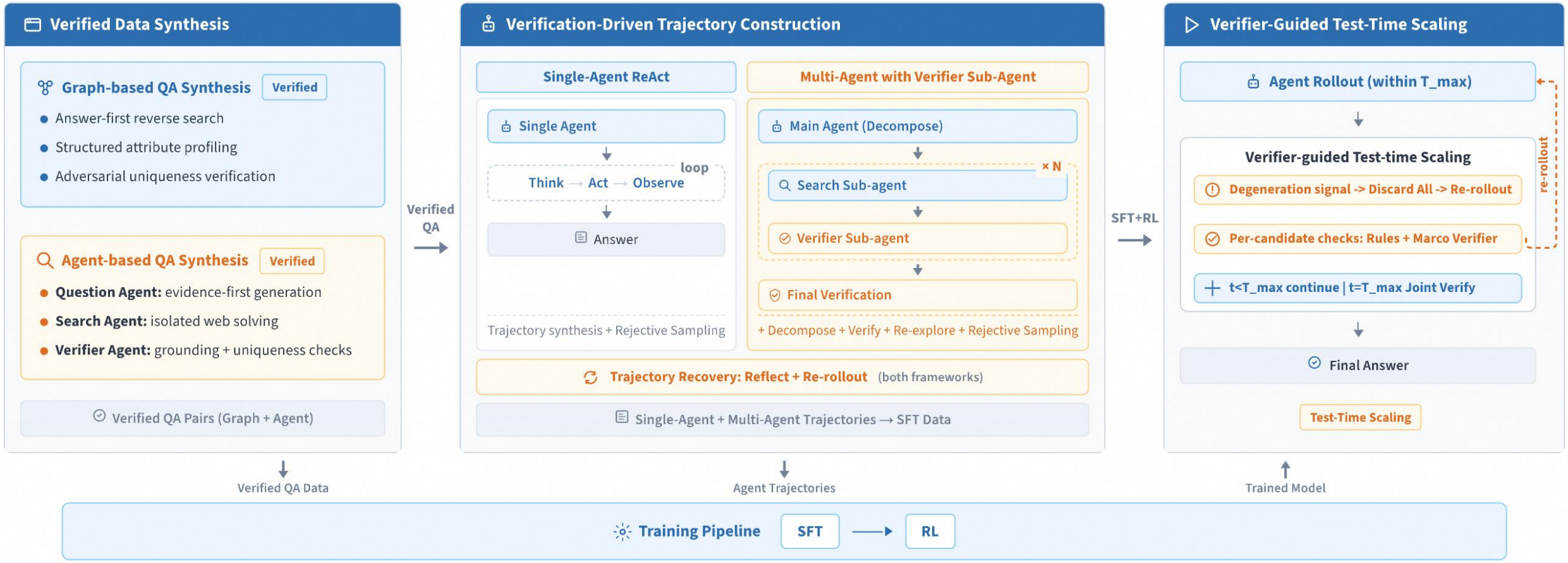
验证驱动的数据合成和TTS

有效的验证机制对长周期Agent非常重要

HSCodeComp	过早进入错误推理路径后难以回溯，需要准确可靠的验证机制
DeepWide Search	长程任务需要依赖规划执行任务且搜索结果需要校验

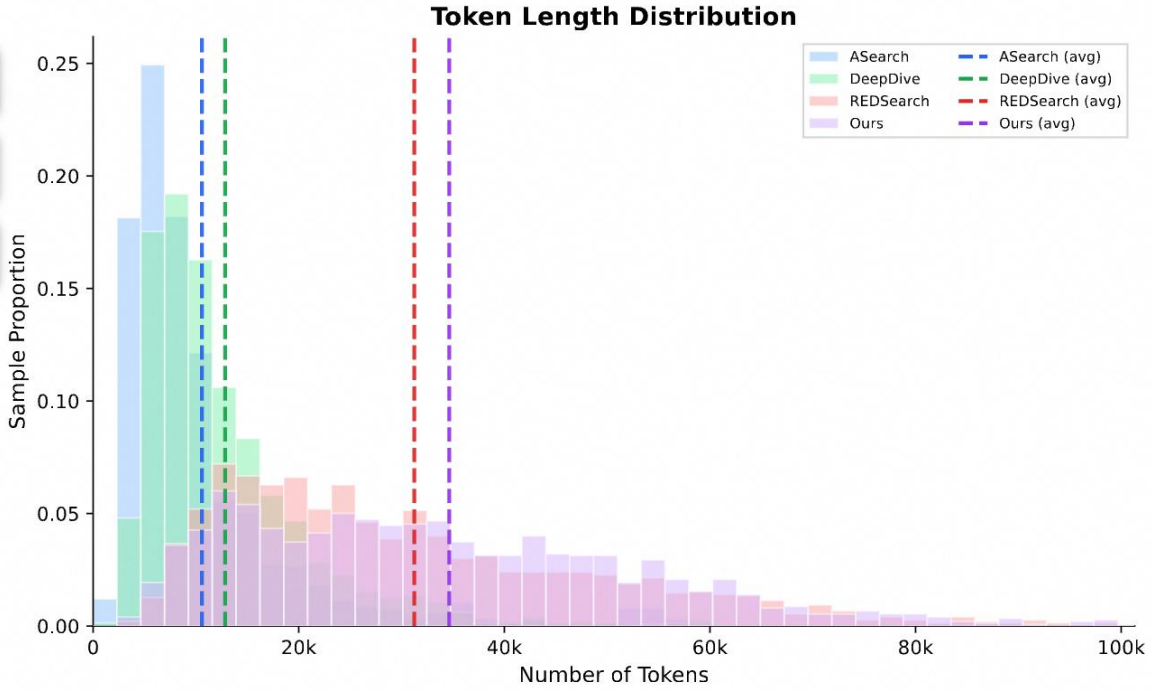
- 当前的Search Agent，尤其是经过Agentic RL优化的Search Agent**缺乏显式的验证机制**
- 错误在数据合成、轨迹合成和推理阶段逐渐累积、影响模型的性能
- 核心在于构建一套包含验证行为的数据合成方法和推理时扩展机制，释放小模型的潜力

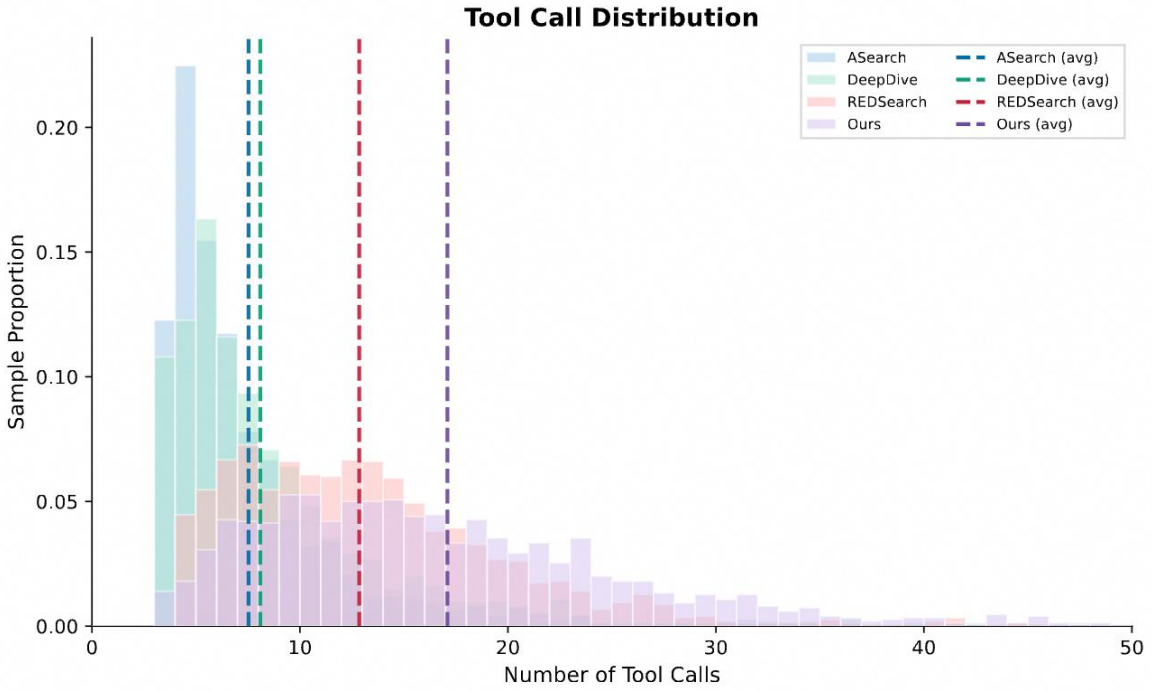
整体框架



从简单堆数据到积累Verification pattern的数据，提升数据合成、轨迹合成的质量和TTS的效果

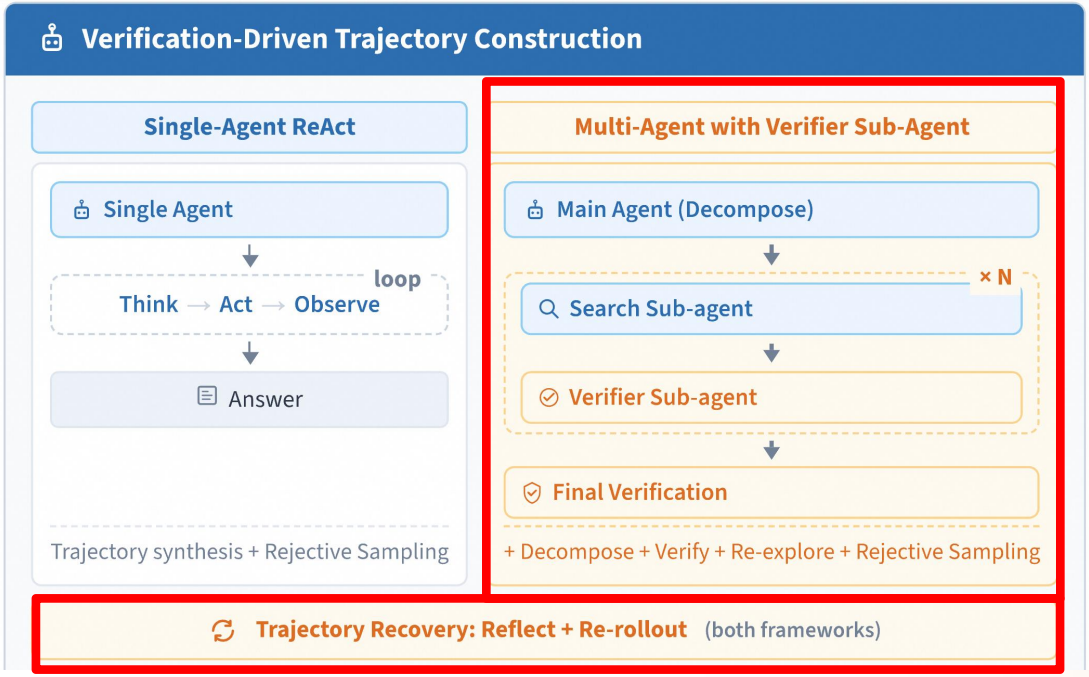
高质量QA数据合成





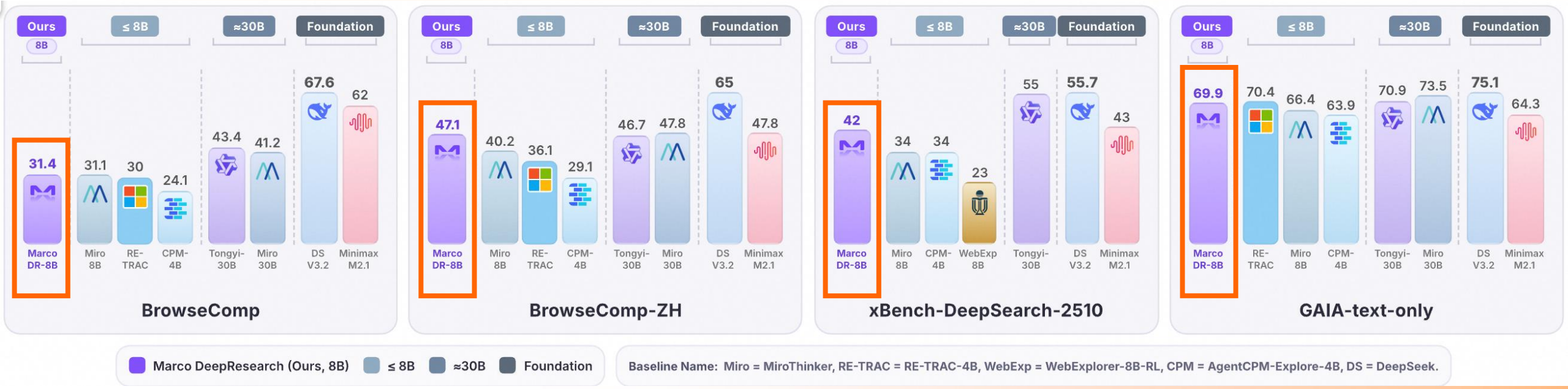
- **核心挑战：**如何自动构造可验证的高质量Multi-hop 深度搜索 QA 数据?
- **多智能体对抗验证闭环：**Generator 生成约束 → Attacker 搜索反例 → Analyzer 补充约束 → 迭代收敛

验证行为轨迹收集 & 基于沙盒的Agentic RL训练



- **SFT冷启动：**收集验证行为轨迹数据，每次子任务和最终答案都需要严格独立的第三方验证
- **基于沙盒的Agentic RL训练的工程优化：**（1）Redis 缓存从根源削减实时 QPS 峰值；（2）分层 LLM-as-a-Judge Reward 计算平衡效率和准确率

主实验结果



- Marco DeepResearch: Qwen3-8B · SFT + GRPO · 128K · 8B 规模达到领先竞争水平
- 超过或匹配部分 30B 级 DeepResearch Agent性能

05 | 真实业务场景应用

Table-as-Search, UMEM, Marco-DeepResearch
在真实业务场景的落地

Table-as-Search 在 AI智能商务拓展（BD）业务场景的应用

Multi-Agent ReAct Baseline

Outcome (Incomplete & Low Precision):

Merchant	Platform	Store Name	Email	Phone	Verdict
Progressive Lighting	Amazon/Site	Lights Online	NA	(866) 688-3562	✓
Brand Name Lighting	Amazon	*Generic Store*	NA	NA	✓
AvitaLights	Etsy	David Avital	NA	NA	✗*(Not US)*
HANM	Etsy	NA	NA	NA	✗*(Invalid)*
Wholesale Lighting	Amazon	NA	NA	NA	✗*(Not Found)*
...	...	...	...	...	...

Performance: Column-F1: 80.0% (False positives), **Item-P: 73.0%** (Missing contacts).

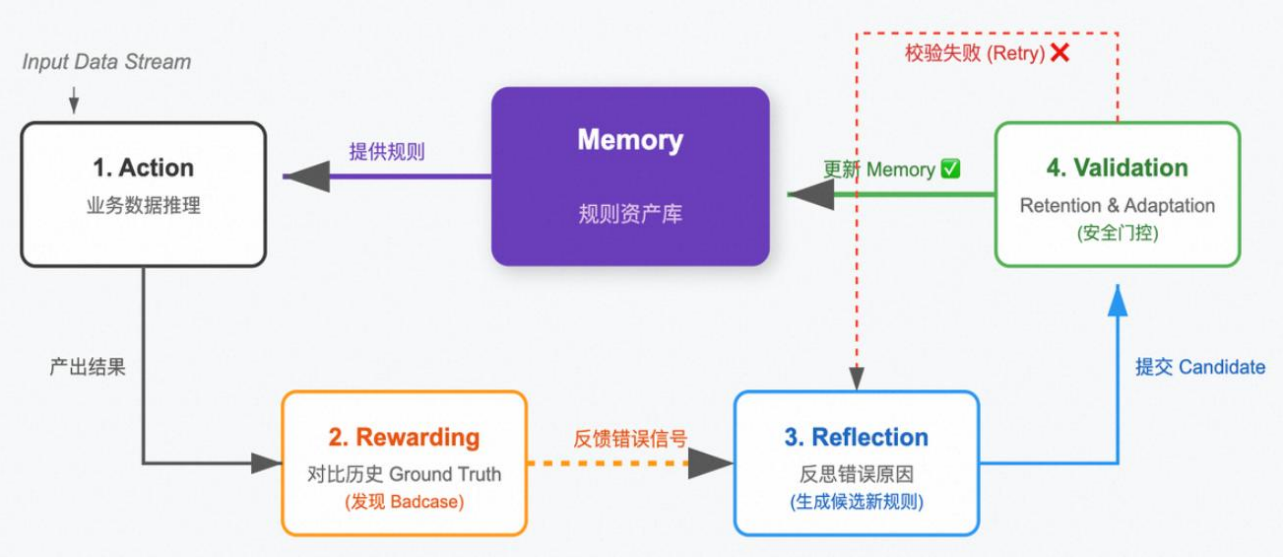
Merchant	Platform	Store Name	Email	Phone	Verdict
Meyda Lighting	Site	Meyda.com	sales@meyda.com	800-222-4009	✓
LFI Lights	Amazon	Light Fixture Ind.	info@lightfixture...	877-534-4621	✓
Hammerton Studio	Site	Hammerton Studio	info@studio.ham...	801-973-8095	✓
HitLights	Amazon	HitLights	customer@hitli...	(855) 768-4135	✓
Commercial LED	Site	U.S. Wholesale	info@commercial...	(313) 528-7900	✓
LightArt	Site	LightArt	info@lightart.com	206-524-2223	✓
...	...	...	...	...	...
TorchStar	Site	TorchStar	info@torchstar.us	(800) 990-7688	✓

Performance: Column-F1: 95.0% (Correctly identified US merchants), **Item-P: 78.9%** (Rich contact details).

Models / Systems	ReAct	Col-F1	Item-P
Gemini DeepResearch	-	51.2	58.3
Claude-Sonnet-4	SA	39.5	35.2
Claude-Sonnet-4	MA	39.3	44.2
Our proposed TaS Framework			
Claude-Sonnet-4 (TaS)	MA	**55.9**	63.5
+ 32B Sub-Agent	MA	52.7	**67.7**

- Claude-Sonnet-4 优于 Gemini-DeepResearch 显著优于商业产品效果
- 替换 Claude-Sonnet-4 为 32B SFT Search Agent基本不影响性能

UMEM 在跨境电商商品审核场景的应用



应用场景 (Application Scenarios)

场景类型	具体任务
透明图检测	白底图审核 (White-background Image Auditing)
情感标题审核	短标题审核 (Short Title Reviews)
品牌商品审核	正品/仿品识别 (Authentic/Counterfeit Detection)

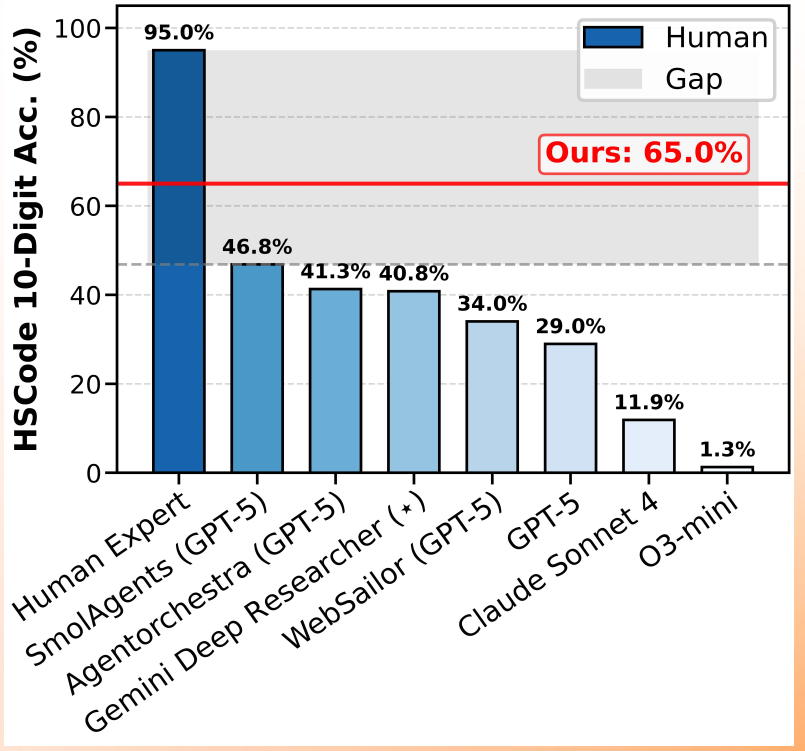
业务效果 (Business Impact)

指标维度	对比基准	UMEM 效果
效率提升	人工调优 3-5 天	10 分钟 自主循环
白底图审核	人工优化基线	+11% 准确率提升
短标题审核	人工优化基线	+2% 准确率提升

大规模电商场景的商品审核场景下，规则往往细微且持续动态变化
人工审核开销大，Self-Evolve Agent Memory 是重要技术解决手段

Agentic RL · Marco DeepResearch在BD业务和HSCodeComp业务应用

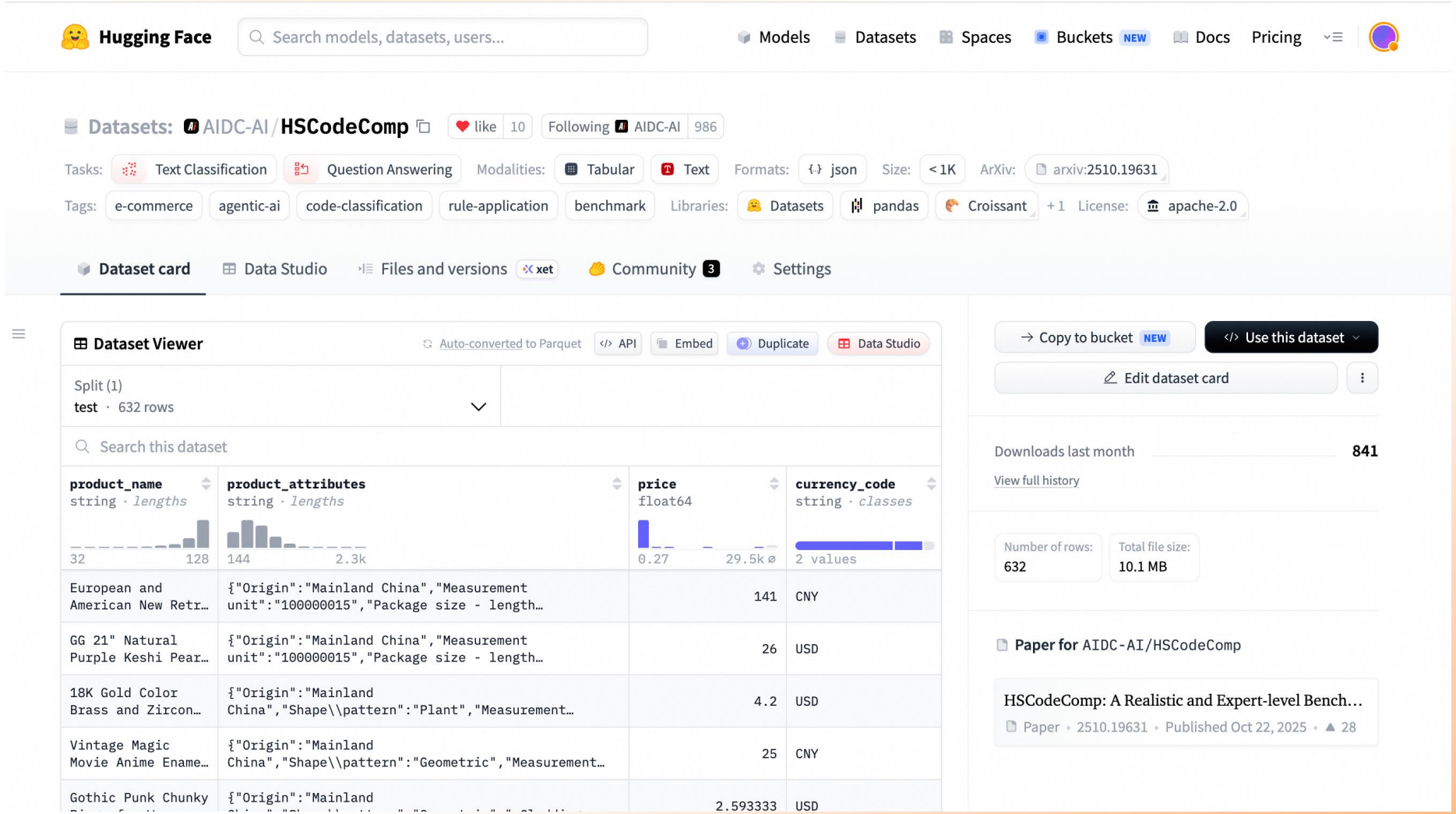
序号	模型 (Model)	模型规模	BD email	GAIA
1	Qwen3-32B-SFT 线上	32B	72	—
3	MarcoDeepResearch 8B-RL Release	8B	68	61.17
6	DR-Venus-4B-RL	4B	46	64.4
7	MarcoDeepResearch 4B-SFT DR-Venus数据 + MarcoDeepResearch数据 + 业务数据	4B	61	66.99



MarcoDeepResearch 在AIBD业务场景：小模型大能量

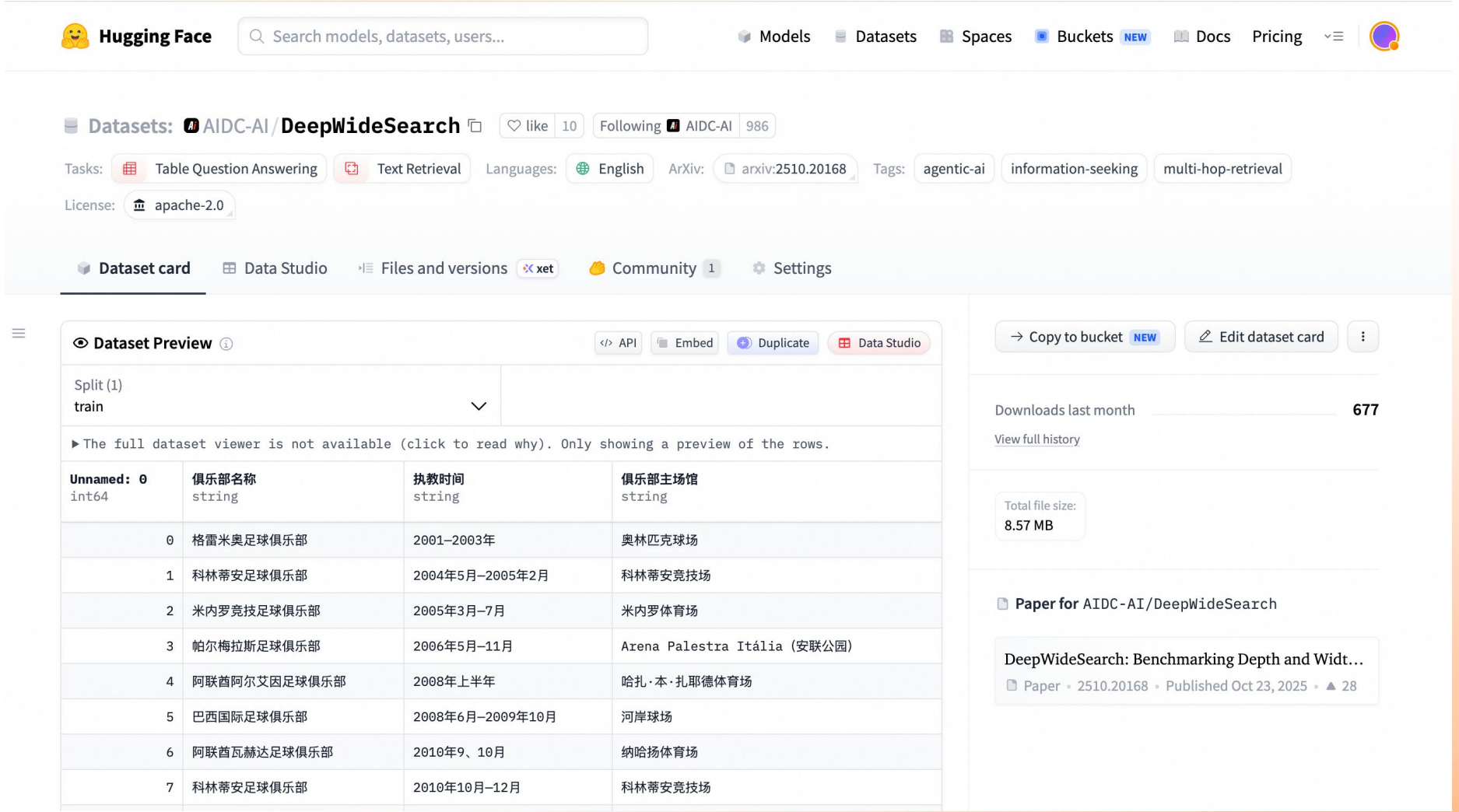
Agentic RL 提升 HSCode
商品分类准确率: +18.2%

技术方案已经开源



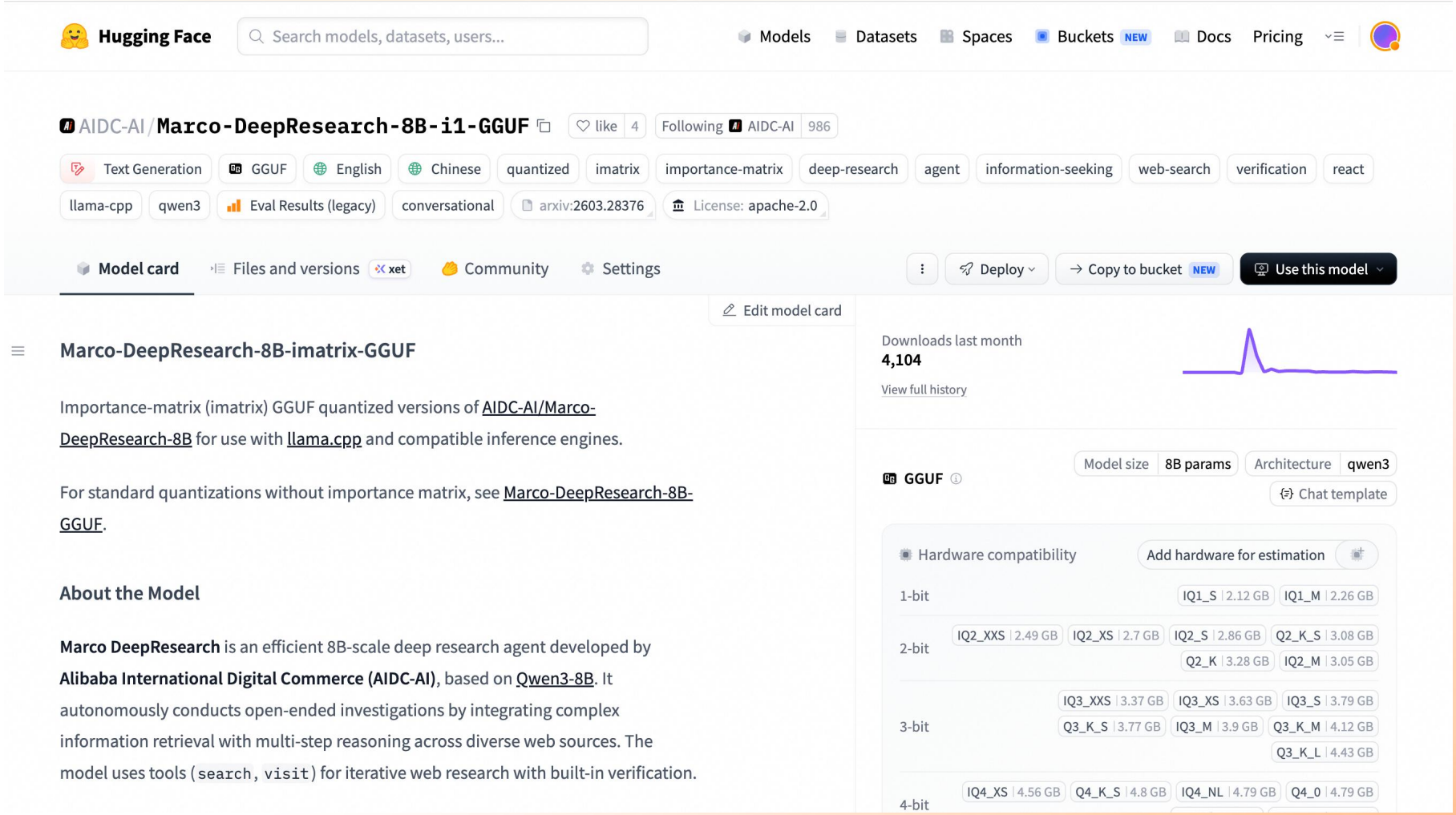
HSCodeComp
下载量 841

技术方案已经开源



DeepWide Search
下载量 667

技术方案已经开源



Marco DeepResearch
下载量 > 4k

Thanks

感谢聆听 · 欢迎批评指正

Contact：lantiangmftby@gmail.com