Agentic RL · 深度研究智能体

LiteResearcher

零边际成本的4B深度检索Agentic RL训练框架

Wanli Li · Bince Qu · Bo Pan · Jianyu Zhang · Zheng Liu · Pan Zhang · Wei Chen · Bo Zhang

浙江大学 · Simplex AI · 香港理工大学

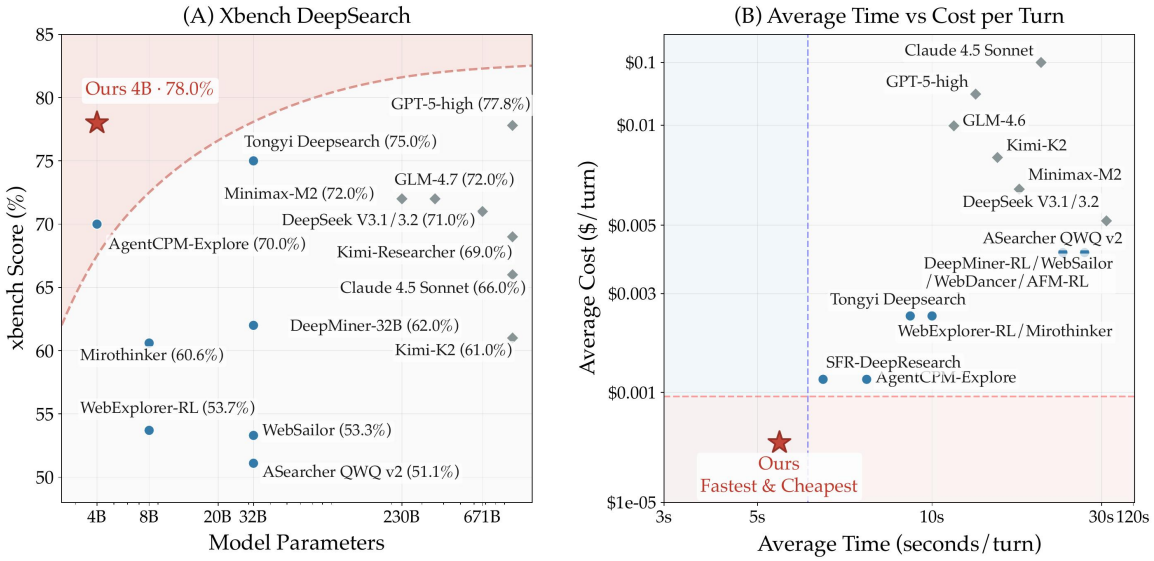
71.3%	78.0%	4B	$0
GAIA	Xbench-DS（开源 SOTA）	参数量	RL 边际成本

github.com/simplex-ai-inc/LiteResearcher huggingface.co/simplex-ai-inc/LiteResearcher-4B

核心结论

一个 4B 模型，以小搏大 8 倍

- **小而 SOTA。**LiteResearcher-4B 媲美甚至超越参数量大其 8 倍的开源智能体与主流商用系统。
- **开源 SOTA。**Xbench-DS 达 78.0%，超过 GPT-5-high（77.8%）与 Tongyi-DR-30B（75.0%）。
- **商用级水平。**GAIA 71.3% ≈ Claude-4.5-Sonnet（71.2%），远超 WebSailor-30B（53.2%）。
- **又快又省。**工具延迟降低 10–46 倍；RL 训练零边际成本。



左：Xbench 准确率 vs 模型规模。　右：每轮延迟与成本 —— 我们最快也最省。

核心论点

对 Agentic 搜索而言，「数据—环境」瓶颈至少与模型规模同等重要。

第一部分 · BACKGROUND

背景与动机

深度研究智能体是什么、Agentic RL 为何难以扩展，以及我们的破局思路。

什么是深度研究智能体？

- **长程信息检索。**拆解开放性问题，联网搜索、阅读网页，综合出有证据支撑的答案。
- **ReAct 范式。**思考 → 行动 → 观察 的循环，往往持续 20–100+ 轮。
- **两个原子动作。**Search(q) 返回排序后的摘要；Browse(u,q) 返回针对查询的页面总结。
- **研究现状。**OpenAI / Gemini Deep Research、Tongyi-DR、WebSailor、Kimi-Researcher —— 多为大模型或闭源。

一次研究举例 *「2008 北京奥运会金牌榜第一的国家？」*

思考	需先定位官方奖牌榜，再读取榜首国家。
行动 · Search	「2008 北京奥运 金牌榜」→ 返回 IOC / 维基等候选。
行动 · Browse	打开维基奖牌榜页 → 抽取榜首行。
观察	页面显示：中国 48 金，居首。
作答	中国 ✓

从「推理 RL」到「Agentic RL」

① 推理 RLVR

- DeepSeek-R1：可验证奖励的 RL 把推理能力内化进权重。
- 涌现出自我验证与回溯等行为。
- 但：封闭世界 —— 仅靠参数化知识，无法与外部交互。

代表工作

DeepSeek-R1 · DAPO · ProRL

② 代码智能体已可扩展

- 隔离沙箱模拟生产环境，过滤基础设施噪声。
- 确定性、低方差的奖励 → 可扩展 RL。
- 证明：环境设计是解锁扩展性的关键。

代表工作

SWE-RL · R2E-Gym

③ 深度研究：尚未实现

- 推理必须与实时工具调用相结合。
- 已有 Agentic RL 难以持续提升。
- 根因：真实网络环境的噪声与不稳定。

代表工作

WebSailor · DeepResearcher · ZeroSearch

关键问题： 既然「隔离沙箱」让代码智能体实现了可扩展 RL，那么深度研究对应的「隔离环境」是什么？

为何深度研究的 Agentic RL 难以扩展

挑战一 —— 数据

- 手工合成数据过度设计推理结构。
- 无法激发真实世界的搜索能力。
- 缺失原子技能：交叉验证、枚举……

挑战二 —— 环境

- 联网 RL → 高方差 + 非确定性奖励。
- 训练规模下 API 成本高昂（$59K–243K）。
- 本地语料 RL → 领域狭窄，无法还原真实网络。

两大挑战相互耦合 —— 共同限制了 Agentic RL 的可扩展性

联网框架（DeepResearcher、WebThinker）须付出方差 + 成本的代价。

本地检索（Search-R1 、ZeroSearch）虽快，却局限于同质化的小语料。

结果：训练饱和 —— 简单题会做、难题全错，模型不再进步。

我们的主张：轻量虚拟世界（孪生架构）

镜像真实互联网的结构 —— 但隔离其执行。

孪生使在虚拟世界中训练的策略可泛化到开放网络；隔离则屏蔽训练过程中的环境噪声。

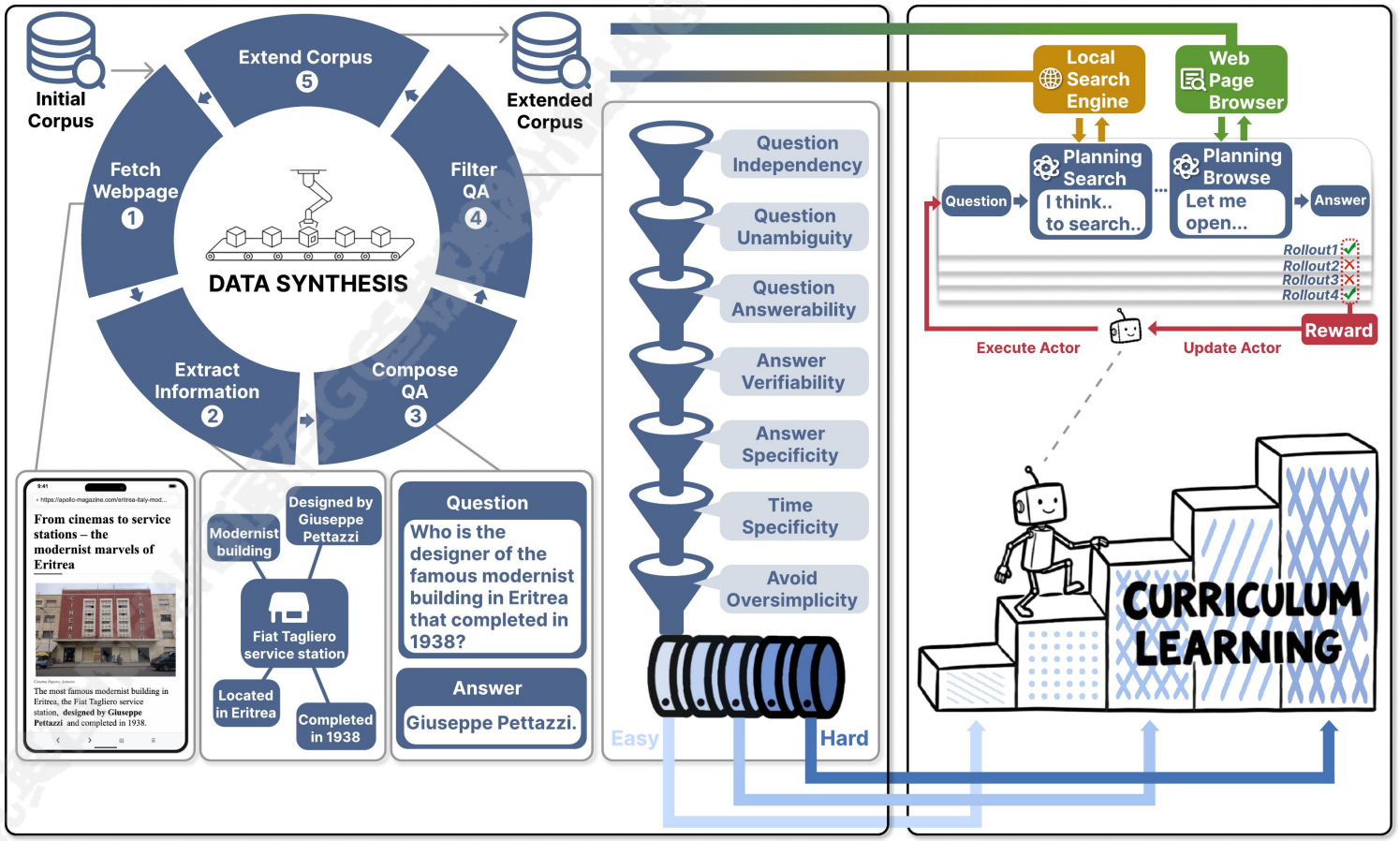
最终效果： *将智能体进化与开放网络解耦，同时保证可迁移回真实网络。*

第二部分 · METHOD

方法：三大支柱

协同构建数据与语料、稳定的本地工具环境、难度感知课程学习。

LiteResearcher 框架的三大支柱



(a) Corpus Extension and QA Synthesis **(b) Reinforcement Curriculum Learning**

(a) 语料扩展与 QA 合成，驱动本地工具。 (b) 强化课程学习。

支柱一

数据 + 语料 协同构建

QA 与本地语料协同进化，兼顾多样性与可扩展性

支柱二

稳定本地工具环境

本地 Search + Browse 镜像真实网络，零成本采样

支柱三

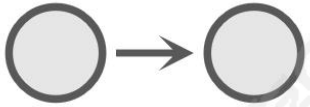
难度感知课程学习

多阶段 RL 始终提供「难度适配」的数据

五种原子搜索能力

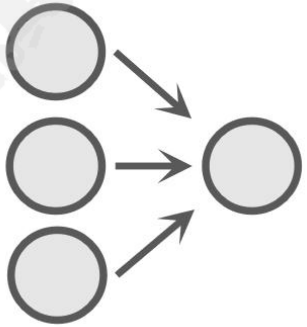
复杂的深度研究轨迹可分解为五种原子能力 —— 我们的数据正是为激发这五种能力而设计。

直接检索



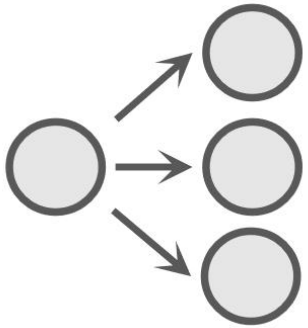
通过搜索定位某一具体事实。

聚合



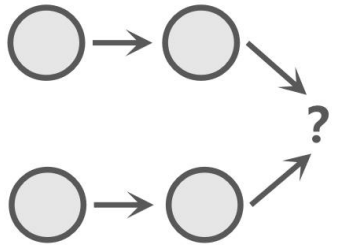
从多个属性约束定位目标 → 取交集。

枚举



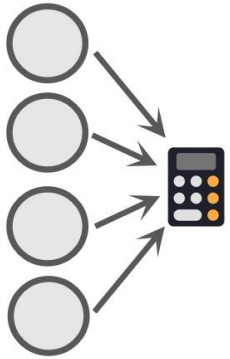
列举并计数满足约束的全部实体 → 取并集。

交叉验证



在多个独立来源间相互印证某一论断。

统计计算



提取数据并计算某个数值指标。

信息源掩码迫使智能体发现非平凡的搜索路径 —— 自然激发出这些能力。

方法 · 支柱一

五种能力 · 合成数据中的真实样例

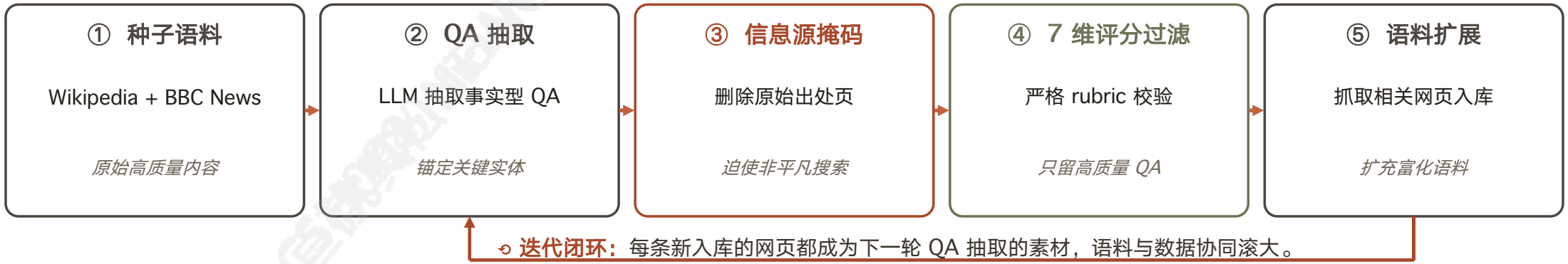
每种能力都对应一类真实可解的合成任务与其「黄金路径」（摘自论文 Table 1）。

能力	合成数据中的样例问题	黄金路径
直接检索	2025 年全球总人口是多少？	① 搜索可靠网页 ② 浏览页面读取数值
聚合	在 2024 年 10 月完成的 Apollo Belvedere 雕像修复项目中，采用传统技法的 Andrea Felice 用什么材料复刻了缺失的左手？	① 归纳全部约束 ② 逐一求解 ③ 取交集
枚举	埃德蒙顿 Valley Line 东南段（市中心至 Mill Woods）于 2023-11-04 开通时共有多少座车站？	① 穷举所有可能来源 ② 逐一访问 ③ 取并集
交叉验证	内蒙古伊利集团旗下「全聪高锌高钙学生奶粉」是否含蔗糖？若含，含量多少？	① 找到多个来源 ② 跨来源交叉验证
统计计算	用于预测南非 O.R. Tambo 区耐药结核（DR-TB）病例的线性回归模型，基于 2018–2021 历史数据，其 R^2 是多少？	① 找到相关数据 ② 计算 R^2 值

要点： *样例均为真实、可验证、需多步检索的难题 —— 不是模板套出来的玩具问题。*

方法 · 支柱一

数据合成流水线：迭代闭环 + 7 维过滤



④ 的 7 维评分准则 *（全部满足才保留，附录 A.1）*

1 问题独立性
脱离上下文也能理解

2 答案具体可验
须是确切数字/名称/日期等

3 问题无歧义
只有唯一解释

4 问题可作答
清楚在问什么

5 非开放式
不用「怎么/为什么」，须短答案

6 非过度简单
常识题剔除

7 时间明确
忌「最新/目前」等模糊时间

设计原则

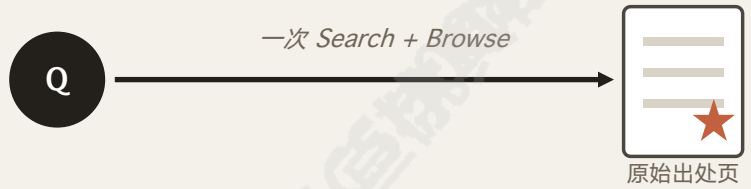
扩大信息源，而非堆砌合成逻辑

- **多样性。**扩大信息源即覆盖真实搜索模式，无需手工推理结构。
- **可扩展性。**简单流水线产出近乎无限的真实任务。

信息源掩码：把「查表」逼成「深度检索」

机制： QA 生成后，从本地语料中删除它的原始出处页 —— 答案仍在语料中，但已无法被一步直达。

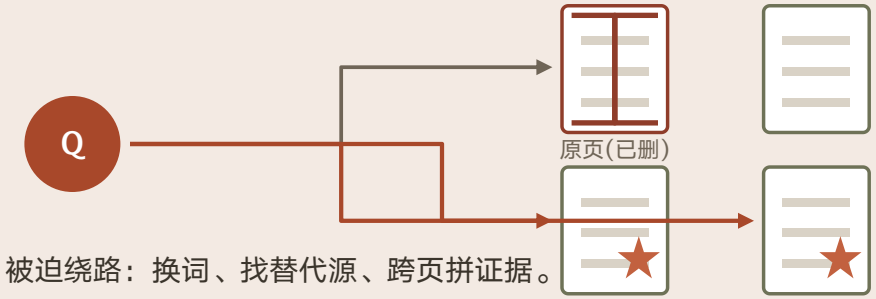
✗ 不做掩码 —— 单跳查表



答案就在原页上 → 一步直达。

学不到任何真实检索技巧。

✓ 删除原始出处 —— 检索多跳



被迫绕路：换词、找替代源、跨页拼证据。

→ 自然涌现 聚合 / 枚举 / 交叉验证。

一个例子（问：某雕像修复用的复刻材料？）

掩码前： 搜「Apollo 修复 材料」→ 命中维基词条 → 直接读出。 **掩码后：** 词条被删 → 改搜修复师 Andrea Felice → 找到新闻稿 → 再交叉博物馆公告确认材料。

为什么这是关键： 难度不靠人工堆砌推理结构，而是由「删源」这一步自动生成 —— 简单、可无限规模化，且贴近真实网络。

多跳合成：反向图推理

- **核心思路。**从图结构「反向」生成问题，使作答必须沿图遍历多跳。

1 · 知识图谱

种子实体 → 搜索 → LLM 抽取特征 → 发现 ≤2 个新实体及关系 → 扩展至 N≤8 节点。

2 · 子图采样

从随机根做 BFS，按边数 + 扰动贪心扩展 → 匿名化的 6 节点子图。

3 · 反向 QA 生成

选定目标实体作为答案 → 将边转为模糊约束 → 组合成自然的 3–5 跳问题。

「信息最小化」设计

- 用模糊指代（如「一位古典晚期作家」）替代具体名称。
- 实体约束最小化 —— 仅够锁定唯一答案。
- 防止被一次简单检索就解出。

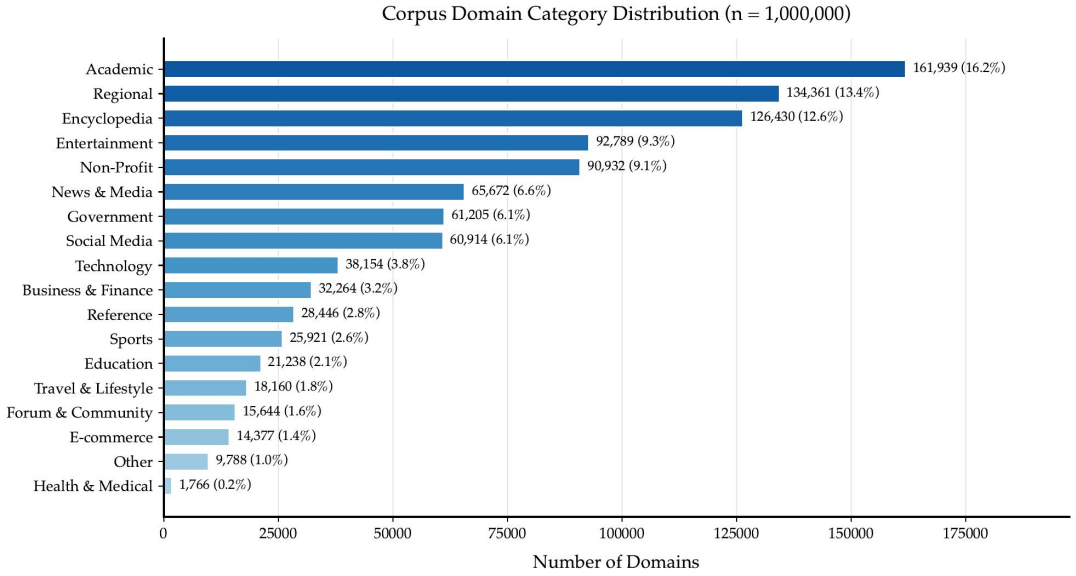
产出

需沿图边进行真正 3–5 跳推理的 QA 对 —— 全部基于真实网络证据。

语料进化：10M → 32M 网页

~32M	1M+	18	$220
网页	独立域名	领域类别	一次性扩展成本

- **迭代增长。**每条通过校验的 QA → 作为搜索查询 → 抓取相关网页 → 追加进语料。
- **10M → 32M。**两轮迭代新增 2200 万真实网页，约 22 万次 Serper 调用。
- **处理流程。**去重（URL + 内容哈希）、HTML→Markdown、>1000 字符过滤。
- **双重索引。**Milvus（搜索）+ PostgreSQL（浏览）—— 即时可供采样使用。



领域分布：学术、地域、百科类来源占据富化语料的最大份额。

一次性 $220 vs. $59K–243K（若 RL 中用在线 API）

方法 · 支柱二

稳定的本地工具环境

在 RL 中取代在线交互 —— 约 3200 万网页全本地服务，支撑数百路并发采样。

本地搜索引擎

- BGE-M3 混合检索（稠密 + 学习型稀疏，一次前向）。
- Milvus 混合融合（RRF）；DiskANN + mmap 落盘。
- 页级索引：每页 1 向量（标题+摘要）—— 索引体积约缩小 10 倍。
- **~0.15 秒/查询** → 比在线搜索快约 10 倍。

本地浏览工具

- 整页内容以 Markdown 存于 PostgreSQL，按 URL 索引。
- 面向 1000 路并发连接调优；URL 上建 B-tree。
- 请求即返回页面内容。
- **~0.17 秒/页** → 比 Jina Reader 快约 46 倍。

吞吐是前提，而非优化项。 整个 RL 过程共 7320 万次工具调用 —— 始终保持零边际成本。

页级（而非分块级）索引让索引足够小，从而支撑数百路并发采样。

方法 · 支柱三

难度感知课程 + on-policy GRPO

- **饱和问题。**当模型被困在某一难度阈值时奖励收敛 —— 简单题会做、难题全错。
- **难度过滤。**每题 K=8 次采样（pass@8）；仅保留 $1 \le c \le 7$ 正确（c=8 过易、c=0 不可解，皆丢弃）。
- **分阶段升级。**每个阶段提升任务难度 + 上下文长度（32K → 48K）。
- **无限数据池。**合成数据的供给让每个阶段都能提供「新鲜且难度适配」的任务。

难度过滤示意：pass@8 保留「会一点但没做透」的题

$c = 0$	8 次全错	✗ 丢弃（不可解 / 太噪声）
$1 \le c \le 7$	部分做对	✓ 保留（有学习信号）
$c = 8$	8 次全对	✗ 丢弃（太简单 / 零梯度）

on-policy GRPO 目标

组相对优势：将每条采样的奖励对组内均值/标准差归一化。

严格 on-policy：每个采样批次只更新一次（mini-batch = batch）。

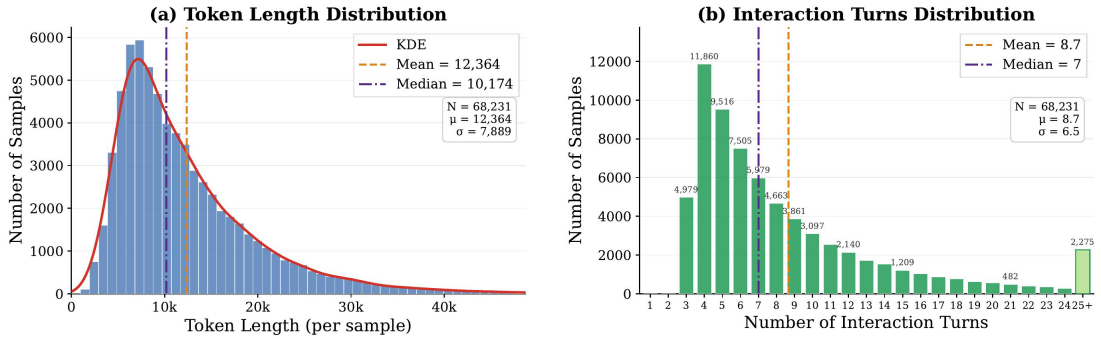
无 KL 惩罚 · 无熵正则。

关键 RL 配置

批量 / K	128 题 × 8 采样
学习率	1e-6 恒定
回复长度	32K（S1）→ 48K（S2）
最大轮数	40（S1）/ 60（S2）
奖励	二值，LLM 判定（Qwen3-30B）
技术栈	VERL · SGLang · FSDP · Ray + TIS

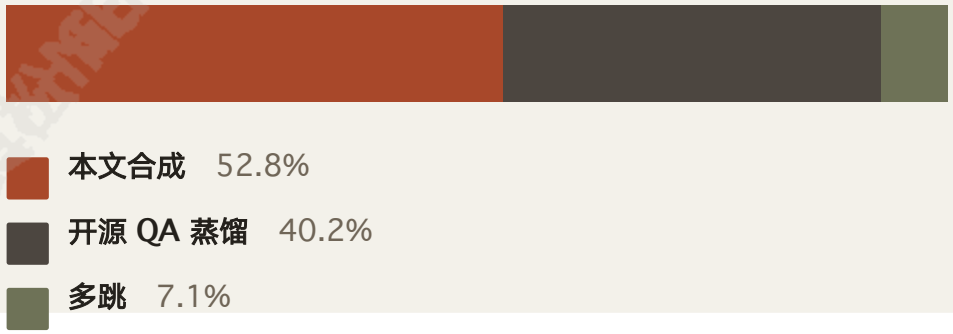
SFT 引导基础工具使用

- **为何先 SFT。**在 RL 之前先教会模型构造查询、发起搜索、浏览网页。
- **教师模型。**Tongyi-DeepResearch 在每条 QA 上用在线工具生成轨迹。
- **68,231 条轨迹。**52.8% 本文合成 · 7.1% 多跳 · 40.2% 开源 QA 蒸馏。
- **清洗。**7 条过滤规则（正确性、无循环、无 Python……）+ 4 步标准化。
- **基座。**Qwen3-4B-Thinking-2507，全参 SFT，64K 序列，LLaMA-Factory / 8×H100。



SFT 数据：平均 12.4K tokens、8.7 轮 —— 长尾促使采用 64K 最大长度。

68,231 条 SFT 轨迹 · 数据构成



GAIA 性能演进

28.2%	55.6%	71.3%
基座	→ SFT	→ RL

RL 再增 +15.7 —— 真正的驱动是训练框架，而非教师蒸馏。

第三部分 · EXPERIMENTS

实验与结果

8 个基准的主结果、成本对比、关键消融，以及训练动态分析。

完整主结果：8 个基准 × 20 个模型

模型	GAIA	BrCmp	Br·ZH	HLE	Frames	WWalk	Seal0	Xben
商用模型								
Claude-4-Sonnet	68.3	12.2	29.1	20.3	80.7	61.7	—	64.6
Claude-4.5-Sonnet	71.2	19.6	40.8	24.5	85.0	—	53.4	66.0
DeepSeek-V3.2	63.5	67.6	65.0	40.8	80.2	—	38.5	71.0
DeepSeek-V3.1	63.1	30.0	49.2	29.8	83.7	61.2	—	71.0
Minimax-M2	75.7	44.0	48.5	31.8	—	—	—	72.0
OpenAI-GPT-5-high	76.4	54.9	65.0	35.2	—	—	51.4	77.8
GLM-4.6	71.9	45.1	49.5	30.4	—	—	—	70.0
Kimi-Researcher	—	—	—	26.9	78.8	—	36.0	69.0
Kimi-K2-0905	60.2	7.4	22.2	21.7	58.1	—	25.2	61.0
开源模型								
Mirothinker 8B	66.4	31.1	40.2	21.5	80.6	60.6	40.4	60.6
Tongyi-DR 30B	70.9	43.4	46.7	32.9	90.6	72.2	—	75.0
ASearcher QWQ v2	58.7	—	—	—	74.5	—	—	51.1
WebSailor 30B	53.2	—	—	—	—	—	—	53.3
WebDancer (QwQ)	51.5	3.8	18.0	—	—	47.9	—	38.3
WebExplorer-8B	50.0	15.7	32.0	17.3	75.7	62.7	—	53.7
DeepMiner-32B	58.7	33.5	40.1	—	—	—	—	62.0
AFM-RL-32B	55.3	11.1	—	18.0	—	63.0	—	—
SFR-DeepResearch	66.0	—	—	28.7	82.8	—	—	—
AgentCPM-Explore-4B	63.9	24.1	29.1	19.1	82.7	68.1	40.5	70.0
LiteResearcher-4B	71.3	27.5*	32.5*	22.0	83.1	72.7	41.8	78.0

*Xbench 78.0% 居开源 SOTA、超 GPT-5-high（77.8%） · GAIA 71.3% ≈ Claude-4.5 · * 表示 64K + 记忆机制。*

实验 · 分析

本地环境：是前提，而非优化项

7320万

RL 中的工具调用

10–46×

延迟加速

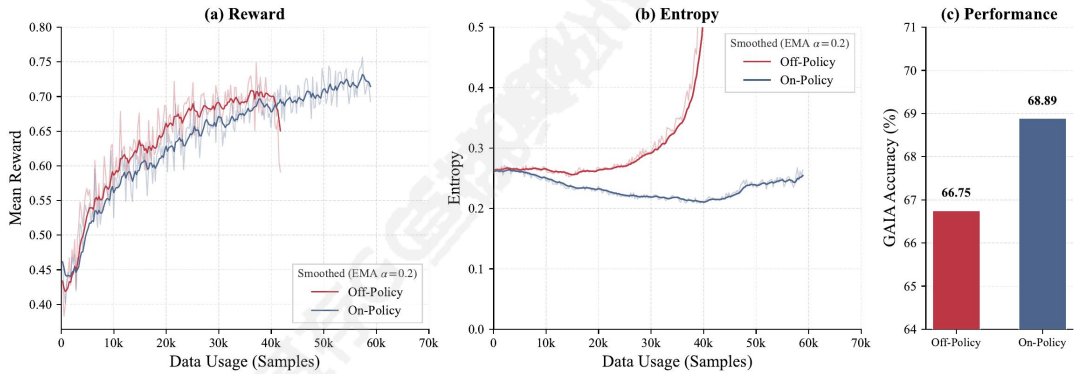
$0

边际成本（vs $59K–243K）

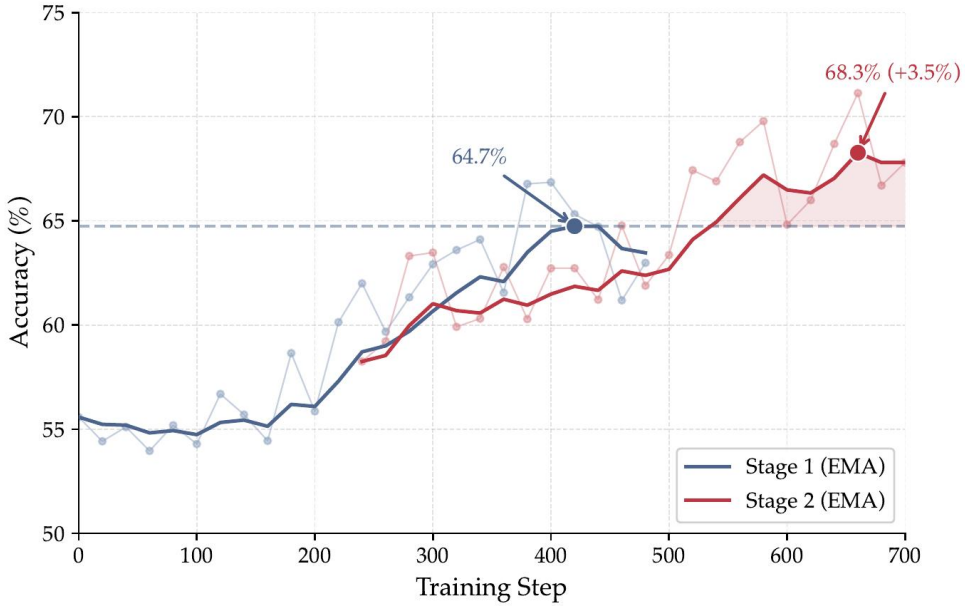
工具 / 提供方	延迟	调用次数	预估成本
Search · Serper	~1.5 s	45.8M	$45,802
Search · SerpAPI	~2 s	45.8M	$229,012
Browse · Jina Reader	7.9 s	27.4M	$13,709
在线合计	—	**73.2M**	**$59K – $243K**
本文（全本地）	**0.15 / 0.17 s**	**73.2M**	**$0**

正是这 10–46× 的吞吐提升，才让 7320 万次调用的 on-policy Agentic RL 成为可能。

on-policy 稳定性 与 两阶段课程



on-policy 持续提升；off-policy 早期涨、随后下滑。



两阶段课程突破 Stage-1 平台期（64.7% → 68.3%）。

- **on-policy > off-policy。**长程 RL 对策略滞后敏感；每批只更新一次 → GAIA 68.9% vs 66.8%。
- **Loss 聚合方式有讲究。**纯 token-mean 下，训练初期超长被截断的零奖励轨迹贡献海量 token、主导梯度 → reward 不涨；改用逐序列等权的 seq-mean-token-mean（标准 GRPO），每条轨迹等权，问题即解。

数据消融：受限环境 + 掩码探索 = 更强的任务能力

核心问题：把我们的合成数据从训练中拿掉，模型在真实基准上会怎样？

训练数据	GAIA	Xbench
仅多跳数据	58.7	66.3
本文数据 + 多跳	66.8	71.0

移除本文数据：

GAIA −8.1 / Xbench −4.7

两类数据都基于真实网页，区别只在合成方式 —— 掉点证明：差距来自我们的「掩码 + 原子能力」设计，而非数据量。

会有人问：多跳不是更难吗？

我们的掩码数据「表面单跳、检索多跳」—— 答案是一个事实，但删源后必须多步绕路才能找到。

一句话：受限 = 把该学的检索技巧主动喂给模型。

为什么受限环境反而更强？

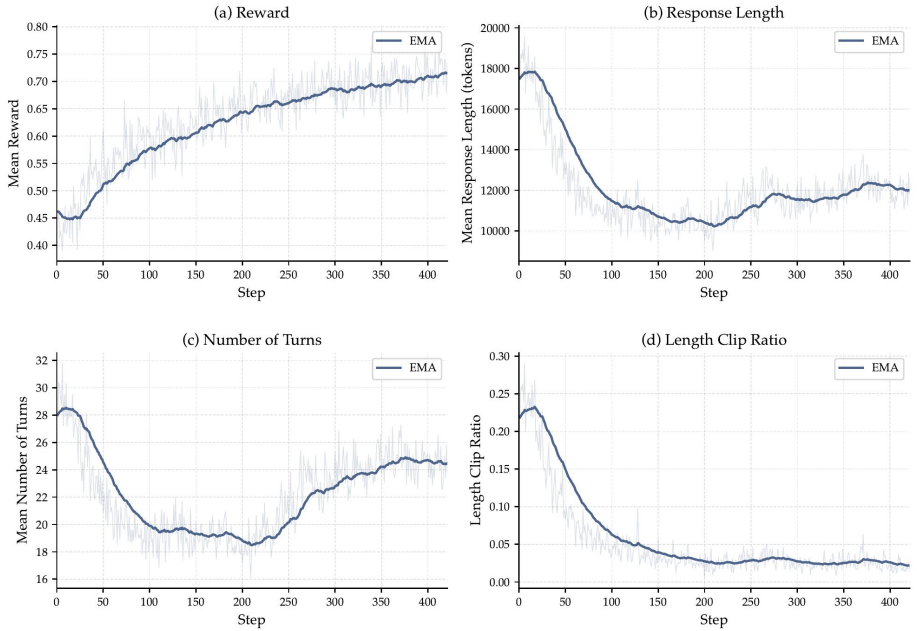
- **掩码逼出主动检索。**删源让一步直达失效，模型必须主动换词、找替代源、跨页取证 —— 正是真实深度检索的核心动作。
- **原子能力被显式激发。**聚合 / 枚举 / 交叉验证 / 统计在合成阶段就被「设计进」任务，而非寄望自发涌现。
- **虚拟世界自洽闭环。**数据、语料、工具同源 —— 任务可解、答案可验，探索能力练在世界里，再迁移到开放网络。

「难」与「对泛化有用」是两个轴

多跳图数据练 **推理深度**；掩码数据练 **检索广度** ——
GAIA / Xbench 考的恰是后者，两者互补、缺一掉点。

RL 消除 SFT 遗留的重复动作循环

- **SFT 失效模式。**重复循环 —— 反复发同一查询/重访同一 URL，白白消耗 token。
- **RL 修正它。**奖励 0.42 → 0.70；回复 18K → 12K tokens；轮数 30 → 24。
- **截断比例。**因超长被截断的轨迹从 0.28 降到 0.02。
- **无显式惩罚。**仅靠结果奖励 + GRPO 裁剪自发涌现。

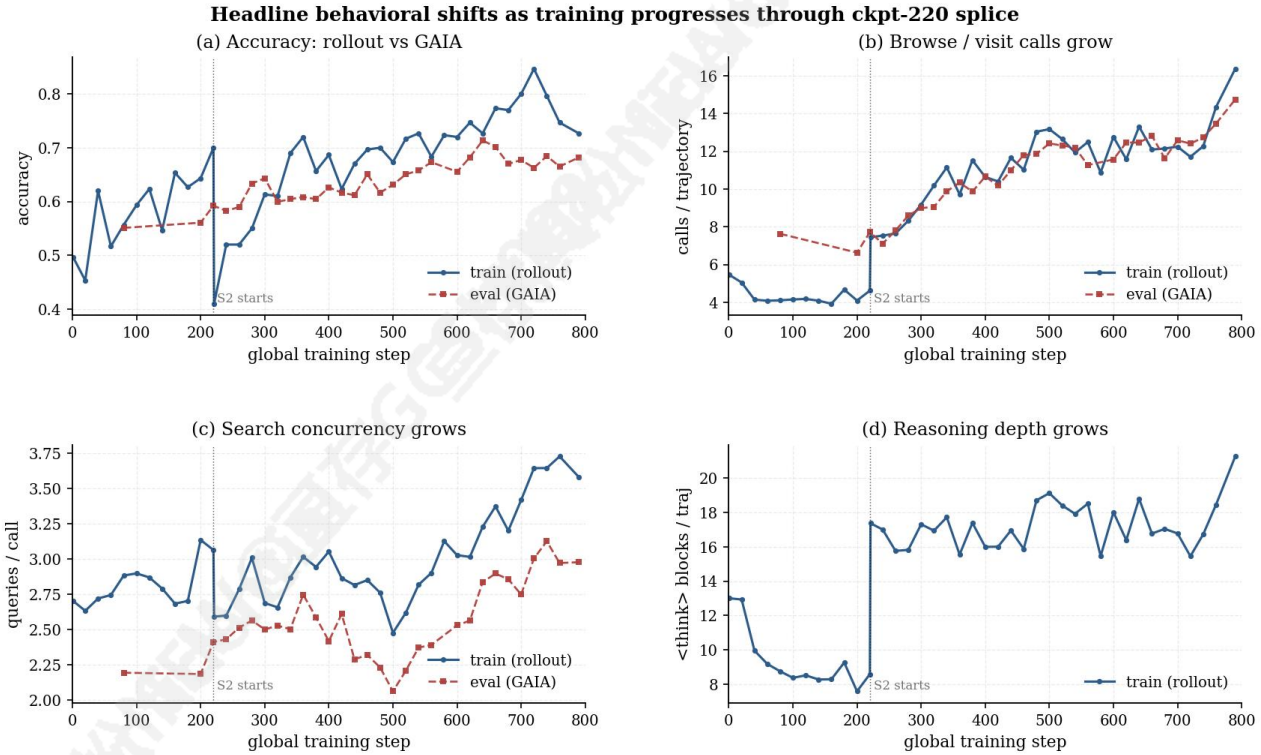


奖励 ↑ 而长度、轮数、截断率 ↓ —— 完全来自结果奖励。

SFT 55.6% → RL 71.3%（GAIA +15.7）

SFT 单独仍落后教师 15.3 分；RL 不仅补齐还反超 —— 8 个基准全面提升。

Sim2Real 收益的四个行为机制 (M1–M4)



41 个 checkpoint × 30+ 行为指标；训练采样（蓝）与 GAIA 评测（红）同步。

四机制同步驱动 GAIA pass@1 +9pt / pass@4 +3pt，主导次序 M1 > M2 > M3 > M4。

M1 工具选择漂移

search → browse：从「再搜一下」变成「再读一页」

browse 比 0.46 → 0.76

M2 轨迹再分配

S1 压缩(假经济) → S2 扩张，多出预算全用于 browse

turns 28→19→45 · visit 4.6→16.4

M3 单次推理变深

think 块频率不变，但每块更长

每块 387 → 626 字 · 占比 9%→14%

M4 自我纠错变多

频率驱动（诚实地说：非新技能）

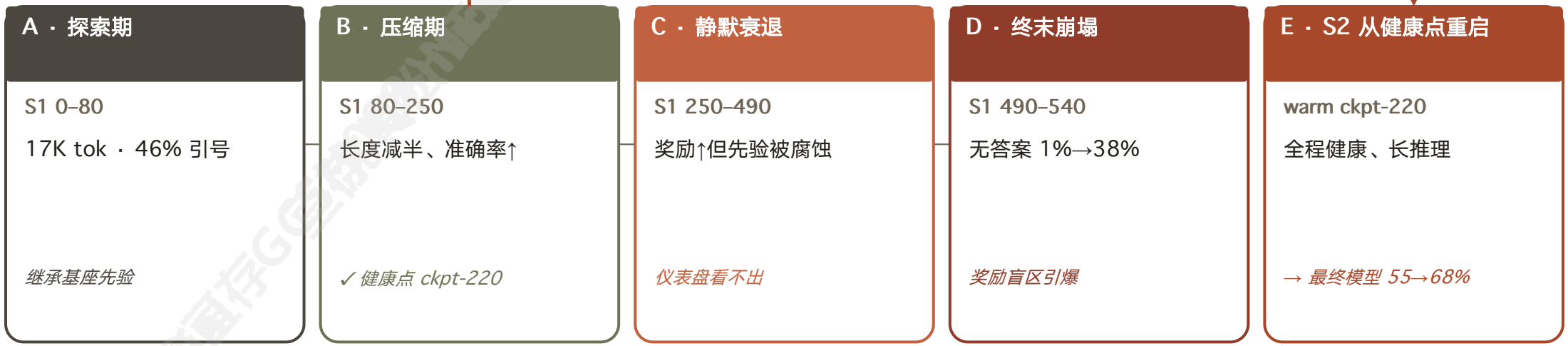
hedge 1.5 → 5.0 /轨迹

一句话：多读页面，每页都仔细想。

为什么必须两阶段：单阶段 GRPO 会确定性崩塌

纯 on-policy GRPO（无 KL、无熵）若一直单阶段训下去，会走完 A→D 直至崩塌 —— 全部发生在 Stage-1。

Stage-2 从崩塌【之前】的健康点 ckpt-220 分叉 —— 非续训崩掉的模型



两阶段不是可选项，而是【绕开】崩塌的必要设计

- **从崩塌前的健康点分叉。**S2 warm-start 自 ckpt-220（崩塌前、准确率≥0.70、无答案≤2%、双语健在），而不是续训已崩的 S1 后期 —— 是 regime reset，不是救援。
- **48K 上下文窗口。**消除扭曲 Stage-1 奖励地形、引爆崩塌的 32K 截断盲区。
- **更难的数据混合。**wiki-16-32K + 多跳子图 + 科学题 —— 走捷径的短轨迹严格更差，从根上消除导致崩塌的压缩力。

第四部分 · CASE STUDIES

案例研究：看智能体怎么想

接下来 5 个真实 rollout，逐轮还原 LiteResearcher-4B 的推理、检索与自我纠错 —— 全部自动判定正确、人工核验无答案泄漏。

DEMO 1 · BrowseComp
大海捞针 · 6 线索

DEMO 2 · Xbench
跨域链式 + 手算日期

DEMO 3 · GAIA
多源核验 + 精确算术

DEMO 4 · Seal-0
精确条件穷尽枚举

DEMO 5 · HLE
第一性原理自推导

案例 · BrowseComp

大海捞针：140 步、6 条零散线索

问： *某位作家曾任大学讲师、作品被译成 20+ 种语言、当过编辑、父母是 NRA 成员、童年常做关于狼的噩梦、有三个孩子（其中两个为收养）——请指出此人。*

第 1 轮	拆解 6 条独立线索；将「狼的噩梦」标记为最稀有的锚点。
第 9–101 轮	排除 20+ 候选：Neal Knox、Carmine Gallo、Stephen King、J.K. Rowling、Baldacci、Sheila Heti……
第 113 轮 · 转向	停止在常见线索上猜测 —— 直接搜「狼的噩梦」这一怪异细节 → Goodreads 访谈。
第 119–138 轮	把六条线索逐一锚定到独立来源：Wikipedia、Lightspeed、Publishers Weekly。
第 140 轮	六条线索全部指向同一人 —— 没有任何单页直接给出答案。

答案：Melissa Marr ✓ —— 70 轮推理、35 次搜索、47 次页面访问；是真正的多跳综合，而非简单查表。

案例 · DEMO 1 · Xbench

跨域链式推理 + 无工具手算日期

问： *语句「如山间清爽的风，如古城温暖的光」出自的小说被拍成电影，该电影上映距中国首枚奥运赛艇金牌已过去多少天？*

第 1 轮	拆解任务链：引文 → 小说 → 电影上映日 → 中国首枚奥运赛艇金牌日 → 两日期相减。先搜引文。
第 2–4 轮	出自张嘉佳《从你的全世界路过》；同名电影，英文维基打不开 → 改用中文维基 → 大陆上映 2016-09-29。
第 5–6 轮	搜「中国第一枚奥运会赛艇金牌」→ sport.gov.cn → 北京 2008 女子四人双桨，注明北京时间 8 月 17 日。
第 8–11 轮	定奖牌日 2008-08-17；识别 2008 为闰年，按年分段累加 → 到 2016-08-17 共 2922 天。
第 13–15 轮	在线计算器给出 2983，与自己的闰年分解不符 → 自检：8×365+2=2922，再加 43 天 = 2965，判定 2983 错误。
第 16–18 轮	多个在线计算器均需交互输入、无法返回数值 → 手算结果成立。

答案：2965 天 ✓ —— 无代码工具，凭闰年规则手算并否决错误的在线计算器。

案例 · DEMO 2 · GAIA

多源核验 + 精确算术

问： *若《Nature》2020 年发表的所有 articles（不含书评/专栏）平均 p 值为 0.04，有多少篇的统计显著性声明会出错？向上取整。*

第 1 轮	需两项：文章数，以及由平均 p=0.04 推出的假阳性率。搜「Nature 2020 文章数」→ 列表页显示 Article (1002)。
第 2–3 轮	访问 nature.com 文章列表确认 1002；但题目限定「仅 articles，不含书评/专栏」，须谨慎归类。
第 4–5 轮	另一列表页却显示 2020 (1037) —— 数字不一致。「articles only」指向 Article 类型过滤 = 1002。
第 6 轮	用 type=article 过滤页明确显示 Article (1002)、2020 (1002) → 确定文章数为 1002。
第 7–9 轮	假阳性部分：平均 p=0.04，自然读作乘积 1002 × 0.04 = 40.08；「向上取整」即天花板 → 41。
第 10–11 轮	排除 1037（research articles）与第三方「12000+」（含全部内容类型，不可靠）。

答案：ceil(1002 × 0.04) = 41 ✓ —— 化解 1002 vs 1037 来源分歧，舍弃不可靠页面，精确取整。

案例 · DEMO 3 · Seal-0

精确条件下的穷尽枚举

问： *有多少位美国前总统在败选普选后再次竞选总统、并在第二次尝试中再次落败？*

第 1 轮	需同时满足 (a) 败选普选 (b) 再次参选 (c) 第二次又败。逐一枚举候选并核验三条件。
第 4 轮	Van Buren —— 1840 败选，1848 以自由土地党再选又败 → 三条件全满足，计入。
第 5 轮	Fillmore —— 1856 以美国党（无知党）参选普选又败 → 计入。
第 6 轮(排除)	Hoover 1932 败选，仅「谋求」1940 提名，从未再成为普选候选人 → 排除。
第 12–17 轮(排除)	T. Roosevelt 此前未曾输过普选；Grant 1880 仅未获提名；W.H. Harrison 死于任内 → 均不符。
第 19–22 轮(交叉验证)	Pew 与 1840/1856 选举页独立确认：仅 Van Buren 与 Fillmore 两人。

答案：2 位 —— Van Buren、Fillmore ✓ —— 按精确措辞排除每一个临界项，穷尽核验后定论。

案例 · DEMO 4 · HLE

网上无解 → 第一性原理自推导

问： *等边三角形 ABC（AB=1），边长 c, 2c, 3c, 4c 的正方形如阶梯般并排置于 BC 上，首/末方块分别接触 AB/BC，求 n=4 时的 c。*

第 1 轮	阶梯式正方形构造，先尝试在线检索这道原题。
第 2–9 轮(检索失败)	YouTube、Stack Exchange、cut-the-knot、AoPS 逐一查；最接近的只有两正方形 a,2a 的帖子 → 决定自行推导。
第 10–11 轮(借鉴方法)	两正方形解法在两端各构成含 60° 的直角三角形并用 tan 60° → 可推广到四方块 c,2c,3c,4c。
第 14 轮(建坐标)	设 B=(0,0), C=(1,0), A=(0.5, √3/2)，则 AB 为 y=√3·x；写出阶梯外角落在 AB 上的接触条件 + BC 总宽约束。
第 15 轮(闭式求解)	两端用 tan 60°=√3，四方块宽 + 两端间隙求和为 1，得关于 c 的线性方程，解得 c = (6−√3)/55。

答案：c = (6−√3) / 55 ✓ —— 网上无现成答案，借鉴双正方形法、建坐标系、闭式自推导。

结论

三点结论

1

Agentic RL 能驱动持续提升

以往工作的饱和源于高方差的在线环境或狭窄语料 —— 而非 RL 范式本身。稳定的虚拟世界可支撑 700+ 步单调提升。

2

简单且可扩展的数据胜过过度设计的逻辑

持续扩展真实世界的信息源、配以简单合成，对搜索模式的覆盖远胜手工设计的推理结构。

3

可扩展 RL —— 而非仅靠模型规模 —— 才是关键

全本地、零边际成本训练，4B 智能体即达 GAIA 71.3% / Xbench 78.0%，媲美甚至超越商用系统。

谢谢

LiteResearcher —— 可扩展 Agentic RL 训练出 4B 深度研究智能体

全栈开源： 数据合成流水线 · 本地环境基础设施 · RL 训练代码

GitHub github.com/simplex-ai-inc/LiteResearcher

模型 huggingface.co/simplex-ai-inc/LiteResearcher-4B

71.3%	78.0%	4B	$0
GAIA	Xbench-DS	参数量	RL 成本

欢迎提问 · 李万里 · wanli_li@zju.edu.cn

联系方式 · CONTACT

李万里 Wanli Li

浙江大学 计算机科学与技术 博士研究生 · CAD & CG 国家重点实验室

研究方向： *Agentic Reinforcement Learning · Multimodal Generation · Large Language Models*

Email	wanli_li@zju.edu.cn
Homepage	个人主页（见 GitHub Profile）
GitHub	github.com/Wanli-Lee
Google Scholar	Wanli Li
微信 / Phone	15560267157

扫码加微信 · 有问题或想交流讨论欢迎联系

项目开源： github.com/simplexai-labs/LiteResearcher · huggingface.co/simplex-ai-inc/LiteResearcher-4B

谢谢观看，欢迎交流合作。