arXiv:2603.07853v1 [cs.AI] 9 Mar 2026

SynPlanResearch-R1：通过合成计划鼓励深度研究中的工具探索

Hansi Zeng [1] **Zoey Li** [2] **Yifan Gao** [2] **Chenwei Zhang** [2] **Xiaoman Pan** [2] **Tao Yang** [2] **Fengran Mo** [3] **Jiacheng Lin** [4] **Xian Li** [2] **Jingbo Shang** [5]

摘要

研究智能体[1] 能够借助工具从网络收集信息，以回答用户查询；这要求它们动态交替进行内部推理与工具使用。尽管原则上可以通过可验证奖励强化学习（RLVR）习得这类能力，但我们观察到，智能体往往表现出较差的探索行为，包括过早终止和工具使用偏向。因此，仅靠 RLVR 带来的提升有限。我们提出 SynPlanResearch-R1，这是一种合成工具使用轨迹的框架，旨在鼓励更深入的探索，以在冷启动监督微调阶段塑造探索行为，并为后续强化学习提供良好的初始化。在七个多跳和开放网络基准测试中，与最先进水平的基线方法相比，SynPlanResearch-R1 在 Qwen3-8B 和 Qwen3-4B 骨干模型上的性能分别最高提升 6.0% 和 5.8%。进一步分析与基线方法相比的工具使用模式和训练动态，揭示了这些提升背后的影响因素。我们的代码已公开发布于 https://github.com/HansiZeng/syn-plan-research。

1. 引言

大语言模型（LLM）在复杂推理、决策以及与外部工具交互方面展现出了卓越的能力（Wang 等人，2025；Feng 等人，2025a；Li 等人，2025d；Dong 等人，2025b）。这推动了研究智能体的发展，它们能够自主探索网络、从各种来源检索并综合信息，并生成全面的答案或报告，以满足用户复杂且知识密集型的查询需求（Nakano 等人，2021；Jin 等人，2025；Li 等人，2025c；Dong 等人，2025c；Feng 等人，2025b）。

研究智能体的前身是检索增强生成（检索增强生成）（Lewis 等人，2020年；Guu 等人，2020年），早期系统直接将文档嵌入输入生成过程，后来的系统则将检索到的文档纳入上下文。一个根本性挑战在于判断何时需要额外上下文，因为人们发现，低质量检索往往会分散模型的注意力，也会降低效率。Self-RAG（Asai 等人，2024年）试图通过训练模型反思检索段落的必要性和相关性来解决这一问题，其奖励由 GPT-4 等强大模型蒸馏而来（Achiam 等人，2023）。然而，这种方法仍局限于非常有限的动作空间：在单轮交互中，模型只能决定是否使用检索，无法控制读取哪些内容或何时停止。时至今日，研究智能体（或深度研究系统）会执行数十轮搜索查询，并配合文档甚至图像处理工具，以应对复杂的用户请求。

我们如何有效训练具备多轮工具使用能力的研究智能体？除了利用专家轨迹进行模仿学习之外，可验证奖励强化学习（RLVR）（DeepSeek-AI 等，2025；Shao 等，2024年；Lambert 等，2025）提供了一种更直接的替代方案。这种在线策略范式仅使用终端任务奖励来优化研究智能体（无论最终答案是否包含真实标注），从而无需训练稠密奖励模型。该过程首先以工具描述和期望行为提示智能体，通常从经由冷启动监督微调（SFT）获得的热启动策略出发，以教会基本的工具使用与格式规范（Jin 等，2025；Li 等，2025c；Dong 等，2025c；b）。随后，该策略仅通过标量结果奖励进行优化。这种优雅的方法使智能体能够凭借自身经验实现自我改进，最终收敛至高性能的任务特定策略。

然而，我们发现，经过可验证奖励强化学习训练的网络研究智能体

*共同贡献 [1]马萨诸塞大学阿默斯特分校 [2]亚马逊 [3]蒙特利尔大学 [4]伊利诺伊大学厄巴纳-香槟分校 [5]圣迭戈大学。通信作者：Hansi Zeng <hzeng@cs.umass.edu>。

[1]研究智能体可以是纯文本的，也可以是基于 GUI 的。本文聚焦于纯文本场景。

往往无法发现解决复杂查询的有效工具使用轨迹。实际中，这种失败通常表现为两种形式：（1）智能体发出的搜索查询太少，过早终止推理；（2）智能体难以在单条推理链中有效组合多种工具，导致证据收集浅尝辄止或支离破碎。我们将这一问题归因于可验证奖励强化学习的同策略特性（Gu 等人，2017）：由于学习过程以智能体自身的 rollout 为基础，薄弱的初始化会限制探索，并使策略陷入次优的工具使用行为。

为解决初始化瓶颈，我们提出 SynPlanResearch-R1，一种在强化学习之前用于冷启动监督微调的计划引导的数据合成框架。其核心思想是合成能够鼓励更深入探索、同时产生正确且格式良好结果的工具使用轨迹，并利用这些轨迹塑造更强的初始策略，从而抬升后续强化学习优化的上限。在 ReAct 框架（Yao 等人，2023）的基础上，我们生成轻量级、随机化的工具计划来引导轨迹生成，促使大型推理模型产生更长的推理链和更丰富的工具交互。为在不抑制模型推理的前提下鼓励遵循计划，我们引入简单的工具相关提示，以柔和方式引导行动选择，并仅保留答案正确且无格式错误的轨迹。在强化学习阶段，我们提出两种简单而有效的稳定策略，以应对多轮基于工具的训练中的挑战，包括对超出轮次或令牌上限的无效轨迹屏蔽策略损失，以及在工具调用违反所需 JSON 模式时立即终止生成，而不是通过反思继续生成。

我们在涵盖多跳问答和开放网络研究任务的七个具有挑战性的基准测试上评估 SynPlanResearch-R1，包括 HotpotQA（Yang 等人，2018）、2WikiMultihopQA（Ho 等人，2020）、MuSiQue（Trivedi 等人，2022）、Bamboogle（Press 等人，2023）、GPQA（Rein 等人，2024）、WebWalkerQA（Wu 等人，2025b）和 GAIA（Mialon 等人，2024）。与以往最先进的搜索智能体相比，SynPlanResearch-R1 在不同模型规模下均取得了稳定提升：使用 8B 模型时，多跳问答的表现最高提升 5.1%，高级问答基准测试提升 8.7%；使用 4B 模型时，分别提升 5.2% 和 6.0%。除了整体性能外，我们还将其与对照基线（例如，基于冷启动监督微调的原始强化学习和从零开始的强化学习）进行分析，发现 SynPlanResearch-R1 能够促进更深入的工具使用探索，并在训练早期保持更高的策略熵，这共同解释了其性能优势。

2. 相关工作

2.1. 可验证奖励强化学习

可验证奖励强化学习（RLVR）（Lambert 等人，2025；Kaufmann 等人，2025）已成为优化大语言模型（LLM）的一种关键范式，应用领域包括数学推理和代码生成（OpenAI，2024；DeepSeek-AI 等人，2025；Shao 等人，2024；Kaufmann 等人，2025）。人们认为，这种方法是 OpenAI o1 系统背后的驱动力（OpenAI，2024），随后又由 DeepSeek-R1（DeepSeek-AI 等人，2025）、Kimi K1.5（Team 等人，2025a）和 QwQ（Team，2024）等模型进一步发展。最近，由于无价值网络优化方法（GRPO（Shao 等人，2024）、RLOO（Ahmadian 等人，2024））效果显著（DeepSeek-R1 已证明这一点），且相较于近端策略优化（PPO）（Schulman 等人，2017）等成熟方法更为简单，因此备受关注。研究者提出了多种方法来提升训练稳定性和效率，包括 DAPO（Yu 等人，2025）、GSPO（Zheng 等人，2025）以及重要性采样校正技术（Yao 等人，2025；Liu 等人，2025a）。尽管可验证奖励强化学习已广泛应用于提升模型性能，但它究竟确实教会了模型新技能，还是仅仅强化了模型已有的能力，仍存在争议（Yue 等人，2025；Liu 等人，2025c）。这与我们的观察相呼应：单靠可验证奖励强化学习无法让模型摆脱自身偏见，从而探索更有效的解决方案。

2.2. 研究智能体

回答知识密集型问题需要与知识库或开放网络进行交互。这一理念推动了检索增强语言模型的发展（Lewis 等，2020；Borgeaud 等，2022；Asai 等，2024；Guan 等，2025；Trivedi 等，2023）。这类模型要么通过扩展语言建模目标进行训练，要么采用非侵入式的多步提示流程。深度研究智能体（OpenAI，2025；Google，2025）通过结合外部工具（尤其是网络搜索和网页浏览），扩展大型推理模型（LRM），以从开放网络中收集、验证和综合证据，回答复杂的知识密集型查询。密集检索可以与语言模型联合训练，但网页浏览行为不可微。近期可验证奖励强化学习的兴起降低了训练研究智能体的门槛（Jin 等人，2025；Wang 等，2025），无需像 Nakano 等（2021）那样收集人工判断来训练奖励模型。当前该方向的研究聚焦于三个互补方面：（1）开发针对基于网页的研究的更有效优化算法（Dong 等，2025c；Jin 等人，2025；Dong 等，2025a）；（2）为冷启动构建高质量训练数据集

监督微调以及后续的强化学习（Sun 等，2025； Li 等，2025c；Tao 等，2025；Li 等，2025a；Liu 等，2025b），以及（3）开发更高效的强化学习框架，以支持长周期训练（Gao 等，2025）。我们的工作旨在通过计划引导的数据合成来增强工具使用能力。

3. 研究智能体学习

3.1. 问题定义

我们将研究智能体定义为一种利用多轮工具调用来响应用户请求的模型。我们采用 ReAct 框架（Yao 等人，2023）来形式化多轮交互场景：每一轮都是一个迭代的**思考**-**行动**-**观察** 循环。在每一步中，模型首先生成内部**思考**，然后执行一项**行动**，最后从环境中接收**观察**。

因此，一个包含 T 轮的轨迹可表示为

$$y^T = (\tau_1, a_1, o_1, \ldots, \tau_T, a_T),$$

其中，τ_i 是第**步**的思考，a_i 是第**行动**，而 o_i 是第**观察**，对应第 i步。模型在 a_T 生成最终答案时，流程终止。

为引导模型遵循这种 ReAct 风格，我们设计了一份简洁的三部分提示词（详见附录 A.1）。它由以下部分组成：(1) 工具描述，列出可用工具，并概述其用途和输入参数；(2)指令，指导模型遵循 ReAct 协议，逐步生成推理、行动和观察；以及 (3) 用户问题，提供实际查询。此外，还加入了一个最简的上下文示例，以提升模型对格式的遵循度。

动作空间除了输出最终答案外，还包含两种外部工具：(i) 网络搜索，用于查询搜索引擎，并返回排名靠前的 k 个候选摘要及其对应的网址；(ii) 网页爬取，用于提取指定网页的内容。

为了在生成的轨迹中明确体现这一行为，我们使用特殊标签包裹每个组件。具体而言，推理词元被包裹在 **<think>、</think>**中，工具操作被包裹在 **<tool call>、</tool call>**中，工具响应则被包裹在 **<toolresponse>、</tool response>**中。当智能体决定生成最终答案时，输出会被包裹在 **<answer>、</answer>**中。这种结构化标注有助于模型内化 ReAct 推理模式，并确保训练和推理过程中的工具使用保持一致。附录8提供了遵循这一格式的智能体输出示例。

3.2. 智能体学习中的挑战

研究智能体可以通过可验证奖励强化学习（Lam- bert 等人，2025；Kaufmann 等人，2025）进行端到端训练。当模型按照第3.1节所规定的带标签 ReAct 结构生成轨迹时，我们可以从 **<answer>、</answer>** 中解析最终答案，并通过与真实标注进行比较，使用完全匹配、F1 重叠度或基于模型评判的准确率来计算奖励。

然而，可验证奖励强化学习训练受到同策略特性所固有的局限。当初始策略初始化不佳或存在偏差时，智能体的探索会受到限制，导致其停滞于次优行为，并收敛到局部最优解。我们发现，训练研究智能体时，这一问题主要表现为两种情况：(1) 智能体倾向于过早结束推理，在尚未执行足够的工具调用以充分探索搜索空间之前，就生成最终答案；(2) 当有多个工具可用时，智能体会强烈偏好熟悉的工具——例如，网络搜索的调用频率远高于网页爬取，从而妨碍对页面内容的深入理解。这种浅层且不均衡的经验动作分布会进一步降低轨迹多样性，并限制后续强化学习更新的效果。

4. SynPlanResearch-R1 框架

我们的整体框架分为两个阶段：先进行冷启动监督微调（SFT），随后使用基于结果的奖励进行强化学习（RL）。为改进初始策略，使其能够进行更深入、更多样的工具使用，我们提出了一种新颖的计划引导的数据生成流程。该流程首先提示大型推理模型（LRM）根据随机化的工具计划生成多条轨迹，以鼓励多样化的工具调用行为。随后，我们应用一组筛选规则，仅保留答案正确且格式有效的轨迹，用于 SFT 训练。

4.1. 计划引导的数据合成

步骤 1：工具计划构建。我们定义了一个函数式工具计划生成器，通过简单但有效的约束来生成合成的工具使用计划。计划长度 L 首先从均匀分布的离散范围 $[L_{\min}, L_{\max}]$中抽取。随后，每个计划都表示为一个序列 P =(a_1、.....、a_L)，其中第一个行动始终设为网络搜索，以便获取候选证据。对于其余步骤（$t \geq 2$），每个行动都从工具集合{网络搜索、网页爬取}中独立抽样，且概率相同（在我们的设定中为 0.5）。

这些计划作为脚手架，鼓励大型推理模型（LRM）生成更丰富的轨迹

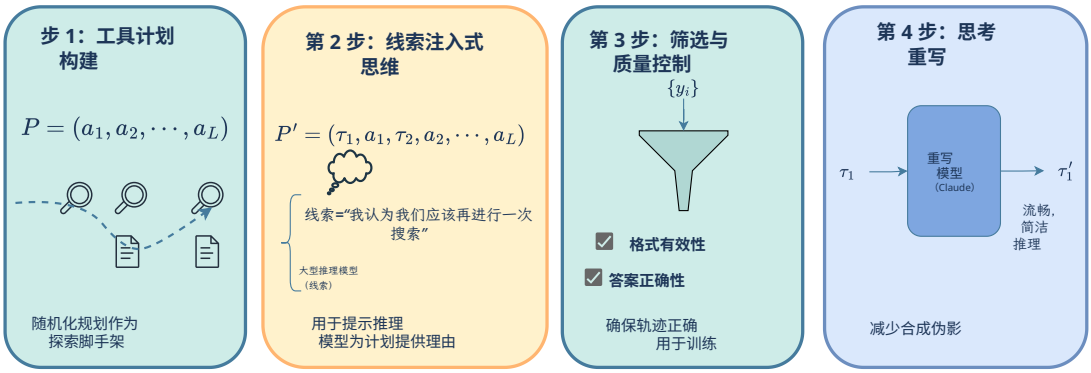

图 *1*。数据合成流程概览：从工具计划生成，到线索注入式多步推理，最后进行轨迹筛选与质量控制。

探索行为，表现为更长的工具使用序列以及更多样化的不同工具组合。

第 2 步：提示注入式思考。生成的工具计划作为辅助指令注入初始用户提示中。然而，这一过程面临两项主要挑战。首先，大型推理模型（大型推理模型）往往无法忠实执行工具计划指令，尤其是在计划指定了多个连续工具调用时。其次，ReAct 框架将轨迹限制为思考-行动-观察循环，要求模型在每次工具调用前生成自然语言思维。这种结构使得难以在行动层面直接落实工具计划。

为了解决这些问题，我们设计了人工编写的提示，并将其注入每个思考步骤的开头。这些提示作为软约束，在保留自然 ReAct 推理流程的同时，温和地引导大型推理模型采取预期的下一步行动。例如，提示可以写道："考虑再进行一次搜索"或"查看其中一个候选网页"。详细的提示模板见附录A.3。

形式化地，给定一个计划 $P = (a_1, \ldots\ldots, a_L)$ 其中包含 $a_t \in${网络搜索、网页爬取}，在第 t 步，我们首先创建一个思考前缀，位于 ReAct 的 `<think>`块内：

$$c_t = \mathrm{Cue}(a_t).$$

然后，我们让大型推理模型从该前缀继续生成思考的剩余部分，随后输出一个行动：

$$(\Delta\tau_t,\ a_t) \sim \mathrm{LRM}(\cdot \mid H_{t-1},\ \texttt{<think>}\ c_t).$$

第 t 步的完整思考由 $\tau_t = c_t \| \Delta\tau_t$给出，其中"‖"表示字符串拼接。

第3步：筛选与质量控制。 对于每个问题，我们从 N 个随机工具计划生成 N 条线索注入的ReAct轨迹 $\{y_i\}_{i=1}^{N}$。若轨迹能够通过这两项检查，我们便予以保留：

- **格式有效性**。 该轨迹符合 ReAct 模式（`<think>`、`<工具调用>`、`<工具响应>`、`<答案>`的顺序和格式正确）。
- **答案正确性**。 最终答案经特定任务检查器验证为正确。

这一过滤过程确保收集到的轨迹既有效又多样——由于工具计划是随机采样的，保留下来的集合自然涵盖了广泛的推理与探索行为。

第 4 步：思维重写。 由于注入的线索较长且由人工构造，可能会引入不自然的措辞或风格偏差。为缓解这一问题，我们遵循（Wu 等人，2025a），采用高质量的改写模型（Claude），在保留其指令意图的同时，将每条线索注入式思维改写为流畅、简洁的语言。这一步减少了合成构造带来的伪影，使生成的轨迹既具有结构化引导，又符合语言自然性，适用于后续的监督微调。思维重写的详细内容见附录 A.4。

4.2. 使用冷启动监督微调进行强化学习

冷启动监督微调。 构建并过滤第 4.1 节所述的合成轨迹后，4.1，我们使用这些轨迹进行冷启动监督微调。从问题集$\mathcal{D}$出发，我们收集相应的合成训练数据，为每个 $\mathcal{Y}_q^+ = \{y_i\}$ 构建一个问题条件集 $q \in \mathcal{D}$。由于合成过程使用的是零提示大推理模型，有些问题无法得到正确的 rollout；我们将这些问题移除。最终得到的监督语料库

为 $\mathcal{D}_{\text{sft}} = \{(q, \ y) \mid q \in \mathcal{D}, \ y \in \mathcal{Y}_q^+\}$，其中每个问题可能包含多条有效轨迹。随后，我们使用该语料库，通过标准最大似然（交叉熵）训练策略 π_θ。最终得到的模型记为 π_{sft}，并作为后续强化学习阶段的初始化。

强化学习。 令 π_{sft} 表示前一阶段得到的冷启动模型。从 π_{sft}开始，我们通过与工具支持环境交互，进一步优化策略。对于每个提示词 $x \in \mathcal{D}$——其中包括工具描述、指令和用户问题——我们使用旧策略 $\pi_{\theta_{\text{old}}}$生成一组 G个 rollout $\{y_i\}_{i=1}^{G}$。每个 rollouty_i 都对应一条 ReAct 风格轨迹，其中涉及工具使用并生成最终答案。

我们使用以下 GRPO 目标函数进行策略优化（Shao 等人，2024年）：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\ \{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\hat{\mathcal{L}}_{\text{GRPO}}(\theta; x, y_i)\right].$$

其中

$$\hat{\mathcal{L}}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \frac{1}{\sum_{t=1}^{|y_i|} m_{i,t}} \sum_{t=1}^{|y_i|} m_{i,t} \cdot \min\Big(\rho_{i,t}(\theta)A_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}\Big).$$

这里，$A_{i,\ t}$ 表示根据每个 rollout y_i的结果奖励计算得到的组归一化相对优势。重要性比率定义为 $\rho_{i,\ t}(\theta) = \frac{\pi_\theta}{(y_{i,\ t}|c_{i,\ t})\ \pi_{\theta_{\text{old}}}(y_{i,\ t}|c_{i,\ t})}$，而$\epsilon$ 是（Schulman 等人，2017）中的裁剪超参数。遵循（Jin 等人，2025年），我们应用二值掩码$m_{i,\ t} \in \{0, 1\}$，其中 $m_{i,\ t} = 0$ 表示该词元来自工具响应，而 $m_{i,\ t} = 1$则表示其他情况。

奖励塑形 结果奖励 $r(q, y_i)$ 根据答案准确率和格式有效性计算。我们使用预测答案与真实答案之间的 F1 分数（记为 s_{ans}）来衡量答案质量。令$f \in \{0, 1\}$ 表示模型输出是否遵循经过验证的 ReAct 格式。奖励定义如下：

$$r(q, y_i) = \begin{cases} s_{\text{ans}} + \alpha, & \text{if } f = 1 \text{ and } s_{\text{ans}} > 0, \\ 0, & \text{if } f = 1 \text{ and } s_{\text{ans}} = 0, \\ 0, & \text{if } f = 0 \text{ and } s_{\text{ans}} \geq \tau, \\ -\alpha, & \text{if } f = 0 \text{ and } s_{\text{ans}} < \tau. \end{cases}$$

其中，α 是与格式相关的奖励系数，$\tau \in (0, 1]$是用于判定何为“好”答案的阈值（在我们的实验中设为 0.8）。

实用训练技巧。 多轮研究智能体经常会产生相当一部分无效轨迹（约 $10\% \sim 15\%$，在Xue 等人，2025）。在这些轨迹中，由于超出最大令牌数或轮次上限，生成过程未能得出有效最终答案。为提升强化学习训练稳定性和样本效率，我们保留这些轨迹用于优势计算，同时通过损失掩码将其排除在策略损失之外。

此外，产生无效工具调用的轨迹——即违反所需 JSON 模式的工具输出——被视为严重失败。在此类情况下，我们会立即终止生成；由此产生的 rollout 不符合所需格式，并被赋予 $-\alpha$的格式相关惩罚。我们在附录C.3中详细说明这两种技巧。

5. 主要结果

5.1. 数据集

评估 我们在四个多跳问答基准测试上评估——HotpotQA（Yang 等人，2018年）、2WikiMulti- hopQA（Ho 等人，2020年）、MuSiQue（Trivedi 等人，2022）和 Bamboogle（Press 等人，2023年）——以及三个可能需要网络搜索的高级推理基准——GPQA（Rein 等人，2024年）、WebWalkerQA（Wu 等人，2025b）和 GAIA（Mialon 等人，2024年）。对于每个数据集，我们使用其官方评估划分（若有测试集则使用测试集，否则使用开发集）。为提高评估效率，我们将 HotpotQA 和 2Wiki-MultihopQA 的评估集下采样至 5K 个实例，而其他所有数据集则使用完整评估划分。

训练 对于监督微调阶段，我们整理了一个包含 8k 个样本的训练集，构成如下：70% 来自多跳问答数据源（HotpotQA、2WikiMultihopQA、MuSiQue 和 MultiHop-RAG（Tang 和 Yang，2024年）），15% 来自 SuperGPQA（Team 等人，2025b），15% 来自 WebWalker-Silver（Wu 等人，2025b）。对于强化学习训练，我们另外构建了一个比例相同的 8k 样本集，并加入 1.6k 个 ARPO（Dong 等人，2025c）样本，使训练实例总数达到约 9.6k（≈10k）。训练集与评估集之间没有重叠。

5.2. 基线方法

我们将 SynPlanResearch-R1 与六种基线方法进行比较。

- 直接推理根据问题直接生成答案，不调用任何外部工具。
- 标准检索增强生成（Lewis 等人，2020）执行单轮检索增强生成，检索排名靠前的k段落，并以此为条件生成模型的回答。
- Search-o1（Liet al.，2025b）是一种多步搜索智能体，它会将查询分解为多个子查询，每个子查询由一个辅助智能体处理，该智能体会交替进行

推理与检索，然后将总结后的证据返回给主智能体，以供后续推理步骤使用。

• 拒绝采样（Yuan 等人，2023）在冷启动阶段使用与 SynPlanResearch-R1 相同的合成数据，但以多轮监督微调和拒绝采样取代强化学习。

• Search-R1（Jin 等人，2025）训练了一个多步搜索智能体，在整个过程中交替进行推理和检索，并通过基于结果的强化学习优化最终答案的准确率。

• SimpleDeepSearcher（Sun 等人，2025）是一种基于强化学习的搜索智能体，着重构建高质量合成数据集，以实现稳健的冷启动初始化。

5.3. 实现细节

我们使用 QWEN3-32B 作为大型推理模型（LRM）来生成合成数据。对于每个问题，工具规划长度 L 从 $[L_{\min}{=}3, L_{\max}{=}8]$中均匀采样，并在每个思考步骤的开头注入一个依赖于工具的提示。每个问题生成$N{=}16$ 个 rollout，我们仅保留同时满足正确 ReAct 格式和最终答案正确的 rollout。

对于冷启动监督微调，我们以 $5e^{-6}$ 的学习率对模型进行 2.5 轮次的微调。对于强化学习训练，我们采用 GRPO 算法，rollout 组大小为 $G{=}16$ ，学习率为 $1e^{-6}$。最大输入长度为 16K 词元，工具调用轮次上限为 8。在强化学习训练的 rollout 阶段，我们使用温度 $=1.0$、top-$p{=}1.0$ 和 top-$k{=}{-}1$。奖励由格式项和正确性项共同构成，格式权重为 $\alpha{=}0.2$，正确性阈值为 $\tau{=}0.8$。

推理时，我们遵循通义千问 Qwen3 模型推荐的默认思考模式解码配置，[2]使用温度 $=0.65$、top-$p{=}0.95$ 和 top-$k{=}20$。词元数量和轮次上限与训练时保持一致。模型性能使用 F1 分数进行评估。

5.4. 性能

表1 比较了 SynPlanResearch-R1 与具有代表性的基线方法在多跳和高级问答基准上的表现。在 8B 规模下，SynPlanResearch-R1 的平均得分达到 0.580，明显优于所有竞争方法。与采用监督微调的拒绝采样（拒绝微调）相比，后者得分为 0.464，我们的方法提升了25。0%。这一提升体现了强化学习更高的采样效率，以及其从正向和负向轨迹中学习的能力，这是迭代式监督微调所不具备的。与 Search-R1 相比，SynPlanResearch-R1 提升了 13。2%，表明冷启动监督微调阶段能够提供更好的策略初始化，从而带来更稳定、高效的强化学习训练。与侧重于过滤高质量问题的 SimpleDeepSearcher 相比，我们通过在冷启动监督微调期间合成更多样化的轨迹，提升了 6。0%。我们认为，这一提升得益于工具规划引导的数据合成，使模型能够接触到更多样化的推理轨迹，而非范围狭窄、经过精心筛选的训练语料库。在 4B 规模下也呈现出类似趋势：SynPlanResearch-R1 的平均表现继续优于拒绝微调（+28. 7%）、Search-R1（+14.5%）和 SimpleDeepSearcher（+5.8%）。

[2]https://huggingface.co/通义千问/Qwen3-8B

6. 分析

6.1. 工具使用与探索

我们衡量了七个基准测试中工具调用平均次数（横轴）与任务表现（纵轴）之间的关系，最后一个面板报告宏平均值。我们将本方法 SynPlanResearch-R1 与两个基线方法进行比较：（i）从朴素冷启动监督微调开始的强化学习；（ii）从零开始的强化学习。对于每种方法，我们通过评估强化学习步数为 20、40、60 和 80 时的检查点来追踪其轨迹，并记录每个检查点的工具调用平均次数及相应得分。

图2 显示，探索深度与性能之间存在明确的关系：工具调用更多的方法始终能取得更高的准确率。随着训练推进，所有方法的工具使用都有所增加，但从零开始的强化学习和从朴素冷启动监督微调开始的强化学习，在早期和最终的探索深度上都受到限制，因此表现不如 SynPlanResearch-R1。这表明，强化学习本身可以优化现有的探索策略，但不足以克服初始化所导致的浅层工具使用先验；相比之下，SynPlanResearch-R1 提供了更强的探索先验，使搜索深度大幅提升，并达到更高的性能上限。

此外，探索预算会根据数据集难度进行调整。对于 GAIA 等较难的基于网络的任务，需要通过更多轮次的证据收集；而经过整理的多跳问答任务则在较少的工具调用后便趋于饱和。这表明，SynPlanResearch-R1 使智能体能够根据任务复杂度动态调整搜索深度。

6.2. 训练动态：熵与奖励

为了进一步理解不同方法的学习行为，我们通过跟踪整个强化学习优化过程中的关键指标来分析其训练动态。图3 展示了三种方法的策略熵（左）和训练奖励（右）的演变：我们提出的 SynPlanResearch-R1、基于朴素冷启动监督微调的强化学习，以及从零开始的强化学习。

表 *1.* 各基准测试上的主要结果。每列中的最佳性能以**粗体**显示。

方法	多跳问答				高级问答			
	HotpotQA	2 Wiki	MuSiQue	Bamboogle	GPQA	WebWalkerQA	GAIA Pass@4	**Avg.**
Qwen3-8B								
直接推理	.270	.296	.114	.175	.184	.140	.146	.189
RAG	.436	.324	.178	.323	.163	.287	.130	.263
Search-o1	.424	.517	.216	.546	.324	.339	.257	.375
拒绝采样	.543	.625	.234	.652	.425	.353	.417	.464
Search-R1	.583	.721	.288	.745	.443	.378	.405	.512
SimpleDeepSearcher	.608	.779	.323	.765	.466	.424	.452	.547
SynPlanResearch-R1	**.639**	**.821**	**.357**	**.784**	**.475**	**.474**	**.510**	**.580**
Qwen3-4B								
直接推理	.229	.300	.091	.111	.182	.179	.119	.173
RAG	.417	.313	.166	.272	.143	.326	.150	.255
Search-o1	.417	.503	.204	.453	.239	.323	.250	.341
拒绝采样	.525	.588	.221	.637	.249	.345	.312	.411
Search-R1	.578	.701	.265	.698	.271	.358	.362	.462
SimpleDeepSearcher	.589	.742	.298	.738	.312	.403	.415	.500
SynPlanResearch-R1	**.611**	**.793**	**.333**	**.753**	**.334**	**.438**	**.439**	**.529**

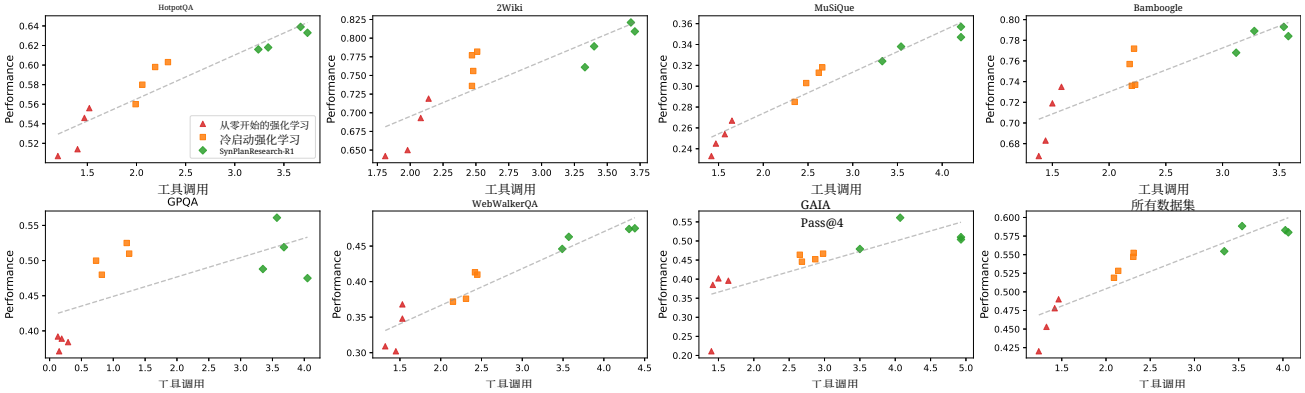

图*2*。七个基准测试中，平均工具调用次数（横轴）与任务表现（纵轴；完全匹配，完全匹配）之间的关系；最后一个面板报告宏平均值。每种方法均展示了四个强化学习检查点（20、40、60 和 80 步）。

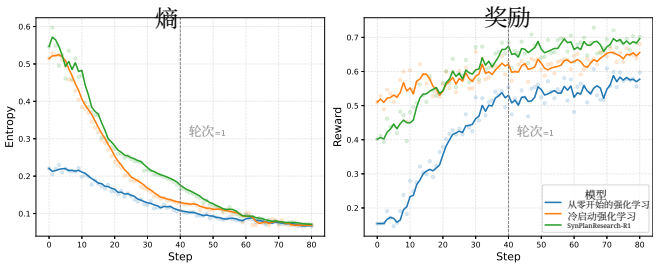

图*3*。三种训练设置下，策略熵（左）和奖励（右）随强化学习步数变化的训练动态。

熵指标揭示了探索能力上的关键差异。我们观察到，与两种基线相比，SynPlanResearch-R1 在整个训练过程中始终保持更高的熵。这种较高的熵表明策略更加多样且更具探索性，这直接源于我们的合成计划引导初始化。在冷启动监督微调期间嵌入的强探索先验使智能体能够对可能的动作保持更广泛的分布，而不是过早收敛到有限的工具使用模式。

奖励动态（图3，右）展现了有趣的学习轨迹。虽然 SynPlanResearch-R1 在训练初期落后于朴素的监督微调+强化学习基线，但它表现出更强的长期进步，最终甚至超越了所有其他方法。这种模式与熵的观察结果一致：更强的探索能力

表 *2*。SynPlanResearch-R1 在各基准测试上的消融研究。每列中的最佳值以**粗体**标出。

方法	MulHop	GPQA	WebWlk.	GAIA@4
SynPlanResearch-R1	**0.650**	0.475	**0.474**	**0.510**
强化学习训练				
1. 最大轮次（8→6）	0.624	0.438	0.456	0.502
2. – 探索展开	0.592	0.414	0.375	0.455
3. – 过滤未完成	0.631	0.457	0.432	0.497
冷启动监督微调				
4. – Claude重写	0.648	0.471	0.461	0.507
5. – 注入思考提示	0.609	**0.511**	0.411	0.472
使用的工具				
6. 仅搜索	0.626	0.374	0.442	0.430

最初会导致行为更具随机性、奖励积累更缓慢，但最终能够发现更优策略，从而取得更高的最终表现。相比之下，朴素的监督微调-强化学习+方法更早进入平台期，这表明由于探索多样性不足，它逃离局部最优的能力有限。

6.3. 消融研究

强化学习训练设置。第*1*行 将工具使用预算上限从 8 步降低到 6 步，导致性能持续下降（例如， −0.26 在多跳问答上， −0.18 在 WebWalkerQA 上），这表明性能与基于工具的探索深度呈正相关。第*2*行将 rollout 阶段的温度从 1.0 降至 0.6，导致各项任务的性能降幅最大（例如， −0.65 在多跳问答上， −0.99 在 WebWalkerQA 上）。这凸显了在强化学习过程中保持足够探索多样性的重要性。第*3*行 将未完成的 rollout（超出 token 或轮次限制）视为负面示例，而非将其过滤掉，这同样导致性能下降（例如， −0.13在 GAIA 上），也反映出训练稳定性降低。

冷启动监督微调组件。第 *4* 行移除了基于 Claude 的思维重写，该方法会将冗长的推理轨迹压缩为更简洁、自然的前缀。它对准确率的影响较小（例如， −0.02 在多跳问答上， −0.03 在 WebWalkerQA 上），这表明性能主要取决于推理能否引导正确的工具使用，而非语言表达是否简洁。相比之下，第 *5* 行移除了合成轨迹生成过程中的提示注入式思考，消除了将模型锚定到目标工具序列的显式逐步指导。这项消融导致四个基准测试中的三个性能大幅下降（多跳问答: −0.31，WebWalkerQA: −0.63，GAIA: −0.38），表明线索在冷启动监督微调期间塑造有效的探索行为方面发挥着核心作用。没有线索时，模型往往会偏离指定的多样化工具计划，导致探索模式更浅并延续到强化学习中。

表3。不同零样本提示方法下，工具调用、网络搜索和网页爬取调用的平均次数，以及生成的数量。Acc 衡量生成的轨迹遵循提示词中指定工具计划的程度。

方法	# 搜索	# 爬取	#总准确率	
标准 ReAct	1.350	0.050	1.400	–
+工具计划	1.480	0.570	2.050	28.32%
+工具计划 +提示	2.720	1.640	4.360	76.96%

工具使用。第 *6* 行将智能体限制为只能使用网页搜索工具。这导致 GAIA 上的表现显著下降（−0.80），表明网页爬取对于复杂的基于网页的推理任务至关重要，尤其是那些需要超越搜索摘要进行更深入证据收集的任务。

6.4. 预训练统计数据

接下来，我们考察不同的零样本提示方案如何塑造用于合成轨迹的大型推理模型的行为。如表3所示，原始 ReAct 提示词只能产生较短且探索性较弱的轨迹：模型平均仅发起 1.35 次工具调用，并且很少执行爬取动作（每个查询 0.05 次调用）。加入明确的工具计划后，大型推理模型会倾向于进行更丰富的探索：网页爬取动作显著增加，工具调用总数也从 1.400 次提升至 2.050 次；但生成的轨迹仅有 28.32% 的时间遵循预设计划，表明仅凭计划不足以实现可靠执行。进一步注入逐步提示后，大型推理模型不仅提高了整体工具使用量（每个查询调用 4.36 次工具，搜索和爬取次数也大幅增加），还在 76.96% 的轨迹中遵循目标计划。这证实了我们的动机：注入提示的计划能够将更强的探索先验印刻到合成数据中，并在强化学习训练期间表现为更高的熵和更深入的工具使用，这一点也与我们对工具使用与性能曲线的分析结果一致。

7. 结论

我们提出了 SynPlanResearch-R1，这是一种计划引导的合成数据框架，可在冷启动监督微调阶段塑造探索行为，从而提升网络研究智能体的能力。在控制下游强化学习、仅改变监督微调初始化的情况下，我们表明，
SynPlanResearch-R1 能够促使智能体进行更深入的工具使用探索，避免朴素初始化中常见的浅层探索和过早终止模式，并持续提升最终性能。在七个具有挑战性的开放领域问答和基于网页的基准测试中，SynPlanResearch-R1 显著优于最先进的基线方法，这凸显了冷启动监督微调是训练稳健网络研究

c智能体的关键杠杆 nts.

影响声明

本文提出了一项方法论贡献，旨在改进对支持多轮工具使用的研究智能体的训练。本研究开发的技术旨在支持通用研究智能体执行信息检索和问答等任务。尽管此类系统可能应用于广泛的下游场景，但这些方法本身并未在现有基于工具的语言模型系统之外引入新的数据收集、用户画像或决策形式。因此，我们认为，除与大语言模型和基于网络的信息检索相关的既有伦理风险外，我们的方法不会带来新的特定伦理风险。总体而言，我们相信这项工作有助于开发更稳健、可控的机器学习系统；除该领域已有的考量之外，我们未发现需要专门讨论的更广泛社会影响。

参考文献

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., 和 Suchir Balaji, e. Gpt-4 技术报告。2023。URLhttps://api.semanticscholar.org/CorpusID:257532815。

Ahmadian,A., Cremer, C., Gall´e, M., Fadaee, M., Kreutzer,¨J., Pietquin, O., Ust¨un, A., and Hooker, S. 回归基础：重新审视用于从大语言模型人类反馈中学习的 Reinforce 风格优化。载于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），第 12248–12267 页，2024年。URLhttps://doi.org/10.18653/v1/2024.acl-long.662。

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG：通过自我反思学习检索、生成和批判。载于第十二届国际学习表征会议，2024年。URLhttps://openreview.net/forum?id=hSyW5go0v8。

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Ruther- ford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., De Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irv- ing, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J., Elsen, E., and Sifre, L. 通过从数万亿词元中检索来改进语言模型。载于 Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. 和 Sabato, S.（编），第 *39* 届国际机器学习会议论文集，第 162 卷，机器学习研究论文集，第 2206–2240 页。PMLR，2022 年 7 月 17–23 日。URL https://proceedings.mlr.press/v162/borgeaud22a.html。

DeepSeek-AI，Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. 等。Deepseek-r1：通过强化学习激励大语言模型的推理能力，2025。Dong, G., Bao, L., Wang, Z., Zhao, K., Li, X., Jin, J., Yang, J., Mao, H., Zhang, F., Gai, K., Zhou, G., Zhu, Y., Wen, J.-R. 和 Dou, Z. 智能体熵平衡策略优化，2025a。URLhttps://arxiv.org/abs/2510.14545。

Dong, G., Chen, Y., Li, X., Jin, J., Qian, H., Zhu, Y., Mao, H., Zhou, G., Dou, Z., 和 Wen, J. Tool-star：通过强化学习赋能大语言模型驱动的多工具推理器。*CoRR*，abs/2505.16410，2025b。数字对象标识符：10.48550/ARXIV.2505.16410。URLhttps://doi.org/10.48550/arXiv.2505.16410。

Dong, G., Mao, H., Ma, K., Bao, L., Chen, Y., Wang, Z., Chen, Z., Du, J., Wang, H., Zhang, F., Zhou, G., Zhu, Y., Wen, J., 和 Dou, Z. 智能体强化策略优化。*CoRR*，abs/2507.19849，2025c。数字对象标识符：10.48550/ARXIV.2507.19849。URLhttps://doi.org/10.48550/arXiv.2507.19849。

Feng, J., Huang, S., Qu, X., Zhang, G., Qin, Y., Zhong, B., Jiang, C., Chi, J., and Zhong, W. ReTool：面向大语言模型策略性工具使用的强化学习，2025a。URLhttps://arxiv.org/abs/2504.11536。

Feng, L., Xue, Z., Liu, T., and An, B. 面向大语言模型智能体训练的组内组策略优化。载于 神经信息处理系统进展，2025b。数字对象标识符：10.48550/arXiv.2505.10978。URLhttps://arxiv.org/abs/2505.10978。

Gao, J., Fu, W., Xie, M., Xu, S., He, C., Mei, Z., Zhu, B., and Wu, Y. 超越十轮：借助大规模异步强化学习，释放长时域智能体搜索的潜力。*arXiv* 预印本 *arXiv:2508.07976*，2025。

Google。Gemini 深度研究——你的个人研究助手。https://gemini.google/overview/deep-research/，2025。访问日期：2025-12-18。

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., Sch¨olkopf, B. 和 Levine, S. 插值策略梯度：融合深度强化学习中的同策略与异策略梯度估计。载于 神经信息处理系统进展，NIPS'17，第 3849–3858 页，美国纽约州雷德胡克，2017 年。Curran

Associates Inc. ISBN 9781510860964。
URLhttps://papers.nips.cc/paper/2017/hash/a1d7311f2a312426d710e1c617fcbc8c-Abstract.html.

Guan, X., Zeng, J., Meng, F., Xin, C., Lu, Y., Lin, H., Han, X., Sun, L., and Zhou, J. Deeprag：大语言模型逐步思考并检索，2025。URL https://arxiv.org/abs/2502.01142。

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. 检索增强语言模型预训练。载于国际机器学习会议，第 3929–3938 页。PMLR，2020。

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. 构建用于全面评估推理步骤的多跳问答数据集。载于 Scott, D., Bel, N., and Zong, C.（编），第 *28* 届计算语言学国际会议论文集，第 6609–6625 页，西班牙巴塞罗那（线上），2020 年 12 月。国际计算语言学委员会。数字对象标识符：10.18653/v1/2020.coling-main.580。URLhttps://aclanthology.org/2020.coling-main.580/。Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., and Han, J. Search-R1：通过强化学习训练大语言模型进行推理并利用搜索引擎。*arXiv* 预印本 *arXiv:2503.09516*，2025。Kaufmann, T., Weng, P., Bengs, V., and H¨ullermeier, E. 基于人类反馈的强化学习综述，2025。URLhttps://openreview.net/forum?id=f70kIurx4b。Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Le Bras, R., Tafjord, Ø., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. T¨ulu 3：推动开放语言模型后训练前沿。在第二届语言建模会议，2025。URLhttps://openreview.net/forum?id=i1uGbfHHpH。Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K¨uttler, H., Lewis, M., Yih, W.-t., Rockt¨aschel, T., Riedel, S., and Kiela, D. 面向知识密集型自然语言处理任务的检索增强生成。载于神经信息处理系统进展，NeurIPS '20，美国纽约州红钩，2020。Curran Associates Inc. ISBN 9781713829546。
URLhttps://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html。

Li, K., Zhang, Z., Yin, H., Zhang, L., Ou, L., Wu, J., Yin, W., Li, B., Tao, Z., Wang, X., Shen, W., Zhang, J., Zhang, D., Wu, X., Jiang, Y., Yan, M., Xie, P., Huang, F., and Zhou, J. Websailor：探索网页智能体的超人类推理能力，2025a。URL https://arxiv.org/abs/2507.02592。Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., and Dou, Z. Search-o1：由智能体搜索增强的大型推理模型。*CoRR*，abs/2501.05366，2025b。数字对象标识符：10.48550/ARXIV.2501.05366。URLhttps://doi.org/10.48550/arXiv.2501.05366。Li, X., Jin, J., Dong, G., Qian, H., Zhu, Y., Wu, Y., Wen, J., and Dou, Z. Webthinker：赋予大型推理模型深度研究能力。载于神经信息处理系统进展，2025c。URLhttps://arxiv.org/abs/2504.21776。Li, X., Zou, H., and Liu, P. Torl：扩展工具集成强化学习，2025d。URLhttps://arxiv.org/abs/2503.23383。Liu, J., Li, Y., Fu, Y., Wang, J., Liu, Q., and Shen, Y. 速度如何损害稳定性：揭示推理与训练不匹配导致强化学习崩溃的原因，2025a。URLhttps://yingru.notion.site/When-Speed-Kills-Stability-Demystifying-RL-Collapse.

Liu, J., Li, Y., Zhang, C., Li, J., Chen, A., Ji, K., Cheng, W., Wu, Z., Du, C., Xu, Q., 等。Webexplorer：探索并演进，以训练长时程网页智能体。*arXiv* 预印本 *arXiv:2509.06501*，2025b。Liu, M., Diao, S., Lu, X., Hu, J., Dong, X., Choi, Y., Kautz, J. 和 Dong, Y. Prorl：延长强化学习拓展大语言模型的推理边界。*arXiv* 预印本 *arXiv:2505.24864*，2025c。
Mialon, G., Fourrier, C., Wolf, T., LeCun, Y. 和 Scialom, T. GAIA：通用人工智能助手基准测试。载于第十二届国际学习表征会议，2024年。
URLhttps://openreview.net/forum?id=fibxvahvs3。
Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B. 和 Schulman, J. WebGPT：结合人类反馈的浏览器辅助问答，2021年。
URLhttps://arxiv.org/abs/2112.09332。OpenAI。学习使用大语言模型进行推理，2024年。URL https://openai.com/index/learning-to-reason-with-llms。

OpenAI。推出深度研究。 https://openai.com/index/introducing-deep-research/, 2025。访问日期：2025-12-18。

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N., and Lewis, M. 衡量并缩小语言模型中的组合性差距。载于 Boua mor, H., Pino, J. 和 Bali, K.（编），计算语言学协会发现：*EMNLP 2023*，第 5687–5711 页，新加坡，2023年12月。计算语言学协会。数字对象标识符：10.18653/v1/2023.findings-e mnlp. 378。URL: h ttps://aclanthology.org/2023.findings-emnlp.378/。

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J. 和 Bowman, S. R. GPQA：研究生级谷歌无法解答的问答基准。载于第一届语言建模会议，2024年。URL：https://openreview.net/forum?id= Ti67584b98。

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., andKlimov, O. 近端策略优化算法，2017年。URLhttps://arxiv.org/abs/1707.06347。

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath：推动开放语言模型数学推理的极限，2024年。

Sun, S., Song, H., Wang, Y., Ren, R., Jiang, J., Zhang, J., Bai, F., Deng, J., Zhao, W. X., Liu, Z., and Wen, J.-R. Simpledeepsearcher：通过网页驱动的推理轨迹合成实现深度信息检索。载于计算语言学协会论文集：*EMNLP 2025*，2025年。 U RLhttps://aclanthology.org/2025.findings-emnlp.739/。

Tang, Y. 和 Yang, Y. Multihop-rag：针对多跳查询的检索增强生成基准。在第一届语言建模会议，2024年。URLhttps://openreview.net/forum?id= t4eB3zYWBK。

Tao, Z.、Wu, J.、Yin, W.、Zhang, J.、Li, B.、Shen, H.、Li, K.、Zhang, L.、Wang, X.、Jiang, Y.、Xie, P.、Huang, F. 和 Zhou, J. Webshaper：通过信息检索形式化进行智能体式数据合成，2025年。URL https://arxiv.org/abs/2507.15061。

Team, K.、Du, A.、Gao, B.、Xing, B.、Jiang, C.、Chen, C.、Li, C.、Xiao, C.、Du, C.、Liao, C. 等。Kimi k1. 5：使用大语言模型扩展强化学习，2025a。

Team, M.-A.-P., Du, X., Yao, Y., Ma, K., Wang, B., Zheng, T., Zhu, K., Liu,M., Liang, Y., Jin, X., Wei, Z., Zheng, C., Deng, K., Guo, S., Jia, S., Jiang, S., Liao, Y., Li, R., Li, Q., Li, S., Li, Y., Li, Y., Ma, D., Ni, Y., Que, H., Wang, Q., Wen, Z., Wu, S., Xing, T., Xu, M., Yang, Z., Wang, Z. M., Zhou, J., Bai, Y., Bu, X., Cai, C., Chen, L., Chen, Y., Cheng, C., Cheng, T., Ding, K., Huang, S., Huang, Y., Li, Y., Li, Y., Li, Z., Liang, T., Lin, C., Lin, H., Ma, Y., Peng, Z., Peng, Z., Qi, Q., Qiu, S., Qu, X., Tan, Y., Wang, Z., Wang, C., Wang, H., Wang, Y., Wang, Y., Xu, J., Yang, K., Yuan, R., Yue, Y., Zhan, T., Zhang, C., Zhang, J., Zhang, X., Zhang, X., Zhang, Y., Zhao, Y., Zheng, X., Zhong, C., Gao, Y., Li, Z., Liu, D., Liu, Q., Liu, T., Ni, S., Peng, J., Qin, Y., Su, W., Wang, G., Wang, S., Yang, J., Yang, M., Cao, M., Yue, X., Zhang, Z., Zhou, W., Liu, J., Lin, Q., Huang, W., and Zhang, G. Supergpqa：在285个研究生学科中扩展大语言模型评测，2025b。 URLhttps://arxiv.org/abs/2502.14739。

Team, Q. Qwq：深入思考未知的边界，2024年11月。URL h ttps://qwenlm.github.io/blog/qwq-32b-preview/。

Trivedi, H., Balasubramanian, N., Khot, T. 和 Sabharwal, A. MuSiQue：通过组合单跳问题构建多跳问题。计算语言学学会会刊，10：539–554，2022。数字对象标识符：10. 1162/tacl a 00475。 URLhttps://aclanthology.org/2022.tacl-1.31/。

Trivedi, H., Balasubramanian, N., Khot, T. 和 Sabharwal, A. 针对知识密集型多步问题，将检索与思维链推理交错进行。载于 Rogers, A.、Boyd-Graber, J. 和 Okazaki, N.（编），计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），第10014–10037页，加拿大多伦多，2023年7月。计算语言学协会。数字对象标识符：10.18653/v1/2023. acl-long.557。 U RLhttps://aclanthology.org/2023.acl-long.557。

Wang, Z., Wang, K., Wang, Q., Zhang, P., Li, L., Yang, Z., Yu, K., Nguyen, M. N., Liu, L., Gottlieb, E., Lam, M., Lu, Y., Cho, K., Wu, J., Li, F.-F., Wang, L., Choi, Y., 和 Li, M. Ragen：通过多轮强化学习理解大语言模型智能体中的自我演化，2025. URLhttps://arxiv.org/abs/2504.20073。

Wu, J., Li, B., Fang, R., Yin, W., Zhang, L., Tao, Z., Zhang, D., Xi, Z., Fu, G., Jiang, Y., Xie, P., Huang, F., 和 Zhou, J. Webdancer：迈向自主信息寻求智能体。载于神经信息处理系统进展，2025a. URL https://openreview.net/forum?id=quJdphBcdP。

Wu, J.、Yin, W.、Jiang, Y.、Wang, Z.、Xi, Z.、Fang, R.、Zhang, L.、He, Y.、Zhou, D.、Xie, P. 和 Huang, F.《WebWalker：大语言模型网页遍历基准测试》。载于 Che, W.、Nabende, J.、Shutova, E. 和 Pilehvar, M. T.

（编），《计算语言学协会第*63*届年会论文集》（第*1*卷：长篇论文），第10290–10305页，奥地利维也纳，2025年7月。计算语言学协会。ISBN 979-8-89176-251-0。数字对象标识符：10.18653/v1/2025.acl-long. 508。URLhttps://aclanthology.org/2025.acl-long.508/。

Xue, Z.、Zheng, L.、Liu, Q.、Li, Y.、Zheng, X.、Ma, Z. 和 An, B. Simpletir：面向多轮工具集成推理的端到端强化学习。*arXiv* 预印本 *arXiv:2509.02479*，2025。

Yang, Z.、Qi, P.、Zhang, S.、Bengio, Y.、Cohen, W.、Salakhutdinov, R. 和 Manning, C. D. HotpotQA：用于多样化、可解释多跳问题回答的数据集。载于 Riloff, E.、Chiang, D.、Hockenmaier, J. 和 Tsujii, J.（编），*2018* 年自然语言处理经验方法会议论文集，第 2369–2380 页，比利时布鲁塞尔，2018 年 10 月至 11 月。计算语言学协会。数字对象标识符：10.18653/v1/D18-1259。URLhttps://aclanthology.org/D18-1259/。Yao, F.、Liu, L.、Zhang, D.、Dong, C.、Shang, J. 和 Gao, J. 高效的 rl 框架暗中带来了离策略强化学习训练，2025 年 8 月。U
R
Lhttps://fengyao.notion.site/off-policy-rl。Yao, S.、Zhao, J.、Yu, D.、Du, N.、Shafran, I.、Narasimhan, K. 和 Cao, Y. ReAct：在语言模型中协同实现推理与行动。载于第十一届国际学习表示会议论文集，2023 年。URLhttps://openreview.net/forum?id=WE_vluYUL-X。Yu, Q.、Zhang, Z.、Zhu, R.、Yuan, Y.、Zuo, X.、Yue, Y.、Dai, W.、Fan, T.、Liu, G.、Liu, L.、Liu, X.、Lin, H.、Lin, Z.、Ma, B.、Sheng, G.、Tong, Y.、Zhang, C.、Zhang, M.、Zhang, W.、Zhu, H.、Zhu, J.、Chen, J.、Chen, J.、Wang, C.、Yu, H.、Song, Y.、Wei, X.、Zhou, H.、Liu, J.、Ma, W.- Y.、Zhang, Y.-Q.、Yan, L.、Qiao, M.、Wu, Y. 和 Wang, M. Dapo：面向大规模强化学习的开源大语言模型强化学习系统，2025 年。URL https://arxiv.org/abs/2503.14476。Yuan, Z.、Yuan, H.、Li, C.、Dong, G.、Lu, K.、Tan, C.、Zhou, C. 和 Zhou, J. 使用大语言模型学习数学推理的缩放关系，2023 年。Yue, Y.、Chen, Z.、Lu, R.、Zhao, A.、Wang, Z.、Yue, Y.、Song, S. 和 Huang, G. 强化学习是否真的能在基础模型之外激励大语言模型的推理能力？载于神经信息处理系统进展，2025 年。URLhttps://openreview.net/forum?id=upehLVgq1b。

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A. 等。组序列策略优化，2025。

A. 提示词

A.1. 研究智能体的 ReAct提示模板

面向研究智能体的 ReAct提示模板

你是一个能够与工具交互以回答复杂问题的智能智能体。以下是可用的工具：

<tool>工具名称：网络搜索 描述：从互联网搜索信息。用法：此工具允许智能体与网络搜索API交互。参数： {"type":"object","properties":"query":"type":"string","required":["query"]}</tool><tool>工具名称：网页爬取 描述：根据网页的 URL 获取网页内容。用法：此工具允许智能体与网页抓取 API 交互。参数： {"type":"object","properties":"url":"type":"string","description":"要获取的网页的 URL","required":["url"]}</tool>

指令：
你以一个或多个循环开始（思考使用哪个工具 → 执行工具调用 → 等待工具响应），并以（思考答案 → 问题的答案）结束。思考过程、工具调用、工具响应和答案均包含在各自的标签内。可能存在多个思考过程、工具调用、工具调用参数和工具响应参数。你应使用的工具为以下之一：网络搜索、网页爬取。
当你到达 <answer> 标签时，你的最终答案必须简短而直接——最好是一个短语或一句话。不要重复问题，也不要再次解释你的推理。推理只应出现在 <think>中。例如，当问题是谁演唱 Beat It 时，你的答案应为 <answer> Michael Jackson </answer>。

示例响应：
<think> 此处为思考过程 </think>
<工具调用>
{{”name”: ”此处为工具名称”, ”parameters”: {”此处为参数名称”: 此处为参数值, ”此处为另一个参数名称”: 此处的另一个参数值，...}}}
</工具调用>
<工具响应>
此处的工具响应
</工具响应>
<思考> 此处为思考过程 </思考>
<工具调用>
{{”名称”: ”此处填写其他工具名称”, ”参数”: {...}}}
</工具调用>
<工具响应>
此处填写工具响应
</工具响应>
（此处还有更多思考过程、工具调用和工具响应）
<think> 此处为思考过程 </think>
<answer> 此处为答案 </answer>

问题： {}

表*4*。用于训练和评估 SynPlanResearch-R1 的完整 ReAct 风格提示词。它包括：(1) 工具描述，(2) ReAct 协议指令，以及 (3) 用户问题。

A.2. 为大型推理模型插入工具计划的 ReAct提示模板

适用于大型推理模型、插入工具计划的 ReAct提示模板

你是 能够与工具交互以回答复杂问题的智能体。以下是可用工具 s:

<工具>
工具名称：网络搜索
描述：搜索互联网中的信息。
用法：此工具可让智能体与网络搜索API交互。
参数： {"type":"object","properties":"query":"type":"string","required":["query"]}
</工具>
<工具>
工具名称：网页爬取
描述：用于根据网页的 URL 获取网页内容的工具。
用法：此工具让智能体与网页抓取 API 交互。
参数： {"type":"object","properties":"url":"type":"string","description":"的 URL 要获取的网页","required":["url"]}
</tool>

指令：
你从一轮或多轮（思考使用哪个工具 → 执行工具调用 → 等待工具响应）开始，并以（思考答案 → 问题的答案）结束。思考过程、工具调用、工具响应和答案都包含在各自的标签内。可能存在多个思考过程、工具调用、工具调用参数和工具响应参数。你可以使用的工具应为以下之一：网络搜索、网页爬取。

以下是你应按顺序遵循的工具使用计划：网络搜索 → 网页爬取 → 网页爬取 → 网络搜索

当你到达 <answer> 标签时，你的最终回答必须简短直接——最好是一个短语或一句话。不要重复问题，也不要再次解释你的推理。推理内容只应放在 <think> 中。例如，如果问题是“谁演唱了《Beat It》？”，你的回答应为 <answer> Michael Jackson </answer>。

示例回复：
<think> 此处为思考过程 </think>
<工具调用>
{{”名称”：”此处填写工具名称”，”参数”： {”此处填写参数名称”：此处填写参数值，”另一个参数名称”：此处填写另一个参数值，...}}}
</工具调用>
<工具响应>
此处填写工具响应
</工具响应>
<think> 思考过程在此 </think>
<工具调用>
{{”name”: ”另一个工具名称在此”, ”arguments”: {...}}}
</工具调用>
<工具响应>
工具响应在此
</工具响应>
（更多思考过程、工具调用和工具响应在此处）
<思考> 思考过程在此处 </思考>
<回答> 回答内容在此处 </回答>

问题： {}

*表5.*一个完整的ReAct提示模板，其中包含用于引导大型推理模型生成多样化轨迹的实例化工具计划。我们将显式工具计划初始化为：网络搜索 → 网页爬取 → 网页爬取 → 网络搜索。请注意，该工具计划是针对每个示例随机生成的。

A.3. 每个行动的注入提示

为了稳定计划条件推理并鼓励大型推理模型忠实遵循规定的工具计划，我们在每个ReAct思考步骤的开头注入简短的提示句。这些提示充当软约束：它们温和地推动模型采取预期的下一步行动，同时保持ReAct式推理的自然语言流畅性。表6 列出了我们针对每种工具行动所使用的确切提示模板。

表 6. 在每个 ReAct <思考> 步骤开始时注入的提示模板。

行动	注入的提示模板
网络搜索	要回答这个问题，我可能需要先收集外部信息。让我想想最好的方式来为网络搜索措辞查询。-
网页爬取	搜索结果中有几个很有希望的链接。也许我应该查看其中一个，以便获取更具体的信息。

A.4. 思维改写

给定一条 ReAct 风格轨迹 $H_L = \{(\tau_t,\ a_t,\ o_t)\}_{t=1}^{L}$，其中 τ_t、a_t 和 o_t 分别表示第 t 步的思考、行动和观察，我们使用 Claude-3.7-Sonnet[3]改写思考。对于每一步 t，我们都会生成一条重写后的思考 $\tilde{\tau}_t$，它 (i) 与原始行动 a_t 和观察 o_t保持一致，且 (ii) 简洁自然。我们按顺序处理轨迹。在第 t 步，改写模型会收到：

- 用户提示 q,
- 先前改写步骤的历史记录 $\{(\tilde{\tau}_i,\ a_i,\ o_i)\}_{i<t}$，以及
- 当前动作 a_t (以及最新观察 o_{t-1}),

并要求其仅输出下一个改写后的思考 $\tilde{\tau}_t$，解释此时采取行动 a_t 为何合理。形式上，改写模型 $\mathrm{LRM}_{\mathrm{rewrite}}$ 的查询方式如下：

$$\tilde{\tau}_t \ = \ \mathrm{LRM}_{\mathrm{rewrite}}\big(q, \{(\tilde{\tau}_i, a_i, o_i)\}_{i<t}, a_t\big),$$

并将轨迹中的原始思考 τ_t 替换为 $\tilde{\tau}_t$。我们对所有步骤 $t = 1$、.....、L重复此过程。第 t 步中用于教师大语言模型的具体提示词见表 7。

第 t步的思考重写提示

你正在模拟一个强大的大语言模型智能体的内部推理过程。
你将看到一个较弱的开源大语言模型智能体尝试分多步完成信息检索任务时的推理轨迹。该智能体以多轮格式与外部工具交互：思考 → 行动 → 观察，目标是回答用户问题。

每一步的格式如下：
思考：..... 行动：
..... 观察：..... —

你的任务是生成当前**思考**的改写版本，称为你的重写思考，简洁自然地解释**当前动作**背后的理由。

你将获得：

- 定义总体目标的**用户提示**,
- 由你的重写思考、行动和观察步骤组成的先前历史记录 ,
- 智能体执行的**当前动作**，以及
- 上一步的最新**观察** .

[3]us.anthropic.claude-3-7-sonnet-20250219-v1:0

指令：

1. 你的重写思考应模拟大语言模型在采取当前行动之前的思考方式。2. 该思考必须从逻辑上论证当前行动的合理性，不得更改或与之矛盾。3. 保持简洁自然；避免过于冗长、风格化或重复的解释。

4. 请勿 重写先前的思考。

5. 仅 输出重写后的思考文本——不要包含任何标签或前缀，例如“你的重写思考：”“思考：_”“行动：”或“观察：”。6. 字面字符串“你的重写思考”不得 出现在输出的任何位置。— 用户提示：“<用户提示>”[迄今为止的轨迹]第 0 步：˜你的重写思考： h_0行动： a_0观察： o_0 ...

第 $t-1$ 步：˜你的重写思考： h_{t-1}行动： a_{t-1}观察： o_{t-1}[轨迹结束]

第 t 步：行动： a_t你的重写思考：

表 7. 第 t步的思维改写提示。

B. 数据合成算法

C. 更多实现细节

C.1. 冷启动监督微调（SFT）。

我们使用多轮 ReAct 轨迹，通过冷启动监督微调（SFT）阶段初始化策略学习。在 8 张 A100 80GB GPU 上采用完全分片数据并行（FSDP）设置，并使用 verl.trainer.fsdp sft训练器进行训练。

我们对基础模型进行 2.5 轮次的微调，学习率为 $5e^{-6}$，全局训练批次大小为 16，每张 GPU 的微批次大小为 4。最大序列长度设为 32,768 个词元。训练期间移除填充令牌以提高效率。为支持长上下文训练，我们采用并行大小为 2 的 Ulysses序列并行。

C.2. 使用 GRPO 进行强化学习。

从冷启动 SFT 检查点开始，我们使用组相对策略优化（GRPO）进一步优化模型，该方法通过 verl.trainer.main ppo 实现。SFT 检查点用于初始化演员和参考模型。训练使用 8 个 GPU，全局批次大小为 256，PPO 小批次大小为 32。演员学习率设为 $1e-6$，训练共运行 80 步。

对于每个输入查询，演员会为每个 GRPO 组采样 $N=16$ 个 rollout。每个 rollout 最多允许 8 个助手轮次，提示长度上限为 16,384 个词元，响应长度上限为 4,096 个词元。rollout 会异步生成，以提升 vllm 引擎支持的吞吐量。在强化学习训练的 rollout 阶段，解码

采用有意随机的方式来鼓励探索，使用温度 $=1.0$、top-p=1.0 和 top-k=-1。在强化学习训练的 rollout 阶段，我们设置温度 $=1.0$、top-p=1.0 和 top-k=-1，以鼓励探索。我们还将 Ulysses 序列并行的大小设为 2，并启用梯度检查点以进行长上下文训练。

奖励函数结合了格式合规性和答案正确性。我们设定格式权重 $\alpha = 0.2$，并使用正确性阈值 $\tau = 0.8$ 来判定生成的答案是否正确。奖励中不包含 KL 正则化（KL 系数设为零），也不进行 critic 预热。

C.3. 强化学习训练稳定性技巧

无效轮次的掩码损失 无效轮次指未能生成有效最终答案的 rollout，仅 因为其超过最大令牌上限或最大交互轮次限制（即被环境截断）。在这种情况下，rollout 不完整，因此无法形成格式良好的最终答案。

令 $x \in \mathcal{D}$ 表示一个提示词，$\{y_i\}_{i=1}^{G}$ 表示在旧策略 $\pi_{\theta_{\text{old}}}(\cdot \mid x)$下采样的一组 rollout。我们将无效轨迹的索引集合定义为

$$\mathcal{U} \triangleq \{\, i \mid y_i \text{ contains a void turn } \}. \tag{1}$$

对于每个 rollout y_i，我们根据第 4.2 节中的奖励定义，计算基于结果的奖励 $r(q,\ y_i)$。对于无效轨迹 $i \in \mathcal{U}$，奖励设为

$$r(q, y_i) = -\alpha, \tag{2}$$

其中 α 是与格式相关的奖励系数。

我们不会让无效轨迹直接更新策略，而是将整条轨迹从策略损失中排除，同时保留它用于组相对优势估计。具体而言，对于所有 $i \in \mathcal{U}$，我们将 rollout 中每个词元的词元级损失掩码设为零：

$$m_{i,t} = 0 \qquad \forall\, t \in \{1, \ldots, |y_i|\}. \tag{3}$$

对于非无效轨迹 $i \notin \mathcal{U}$，掩码遵循标准的 GRPO 设置：对于来自工具响应的词元，设为$m_{i,t} = 0$；$m_{i,t} = 1$ 其他情况。

采用这种轨迹级掩码后，GRPO 目标函数变为

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\ \{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{\sum_{t=1}^{|y_i|} m_{i,t}} \sum_{t=1}^{|y_i|} m_{i,t}\ \min\Big(\rho_{i,t}(\theta)A_{i,t},\ \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}\Big) \right], \tag{4}$$

其中$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|c_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|c_{i,t})}$ 是重要性比率，而$A_{i,t}$ 是根据 rollout 级奖励 $\{r(q, y_j)\}_{j=1}^{G}$计算得出的组归一化相对优势。

由于当 $i \in \mathcal{U}$时，对于所有 t 均有$m_{i,t} = 0$，无效轨迹会参与用于计算$A_{i,t}$ 的组统计量（通过其奖励），但对策略参数不会产生梯度。这样可以防止截断的 rollout 直接扰乱策略更新，同时仍允许它们影响每个 GRPO 组内的相对优势估计。

处理工具调用中的 JSON 模式错误。对于工具使用型智能体，每次工具调用都应遵循预定义的 JSON 模式。在早期实现中，当模型在调用工具时生成无效 JSON 输出，我们会将相应的解析错误（例如 json.loads 引发的错误）作为工具响应返回，并允许生成继续进行。典型的错误消息格式如下：JSONDecodeError: 在位置 42 处应为“,”分隔符。我们发现，这种策略对于通义千问 Qwen3-8B 等较大的模型效果尚可。然而，对于较小的模型（例如通义千问 Qwen3-4B），这种做法会导致严重的训练不稳定性：在大约 40 个强化学习步数后，模型经常会陷入崩溃状态，生成大量格式错误且包含无效 JSON 模式的工具调用，进而导致训练崩溃。为缓解这一问题，我们采用更严格的 JSON 模式违规处理策略。具体而言，对于任何在工具调用期间生成无效 JSON 模式的 rollout，我们会立即终止其轨迹，并将其视为格式违规。

C.4. 网页搜索工具详情

我们使用 Google Serper API 实现网页搜索工具，以获取每个查询的前 10 条搜索结果。Serper 返回的原始搜索结果以 JSON 格式提供，并包含每个 URL 的结构化元数据。我们不会将原始 JSON 输出直接暴露给大语言模型，而是将搜索结果转换为简洁的 Markdown 风格文本表示。每条搜索结果的格式如下：

{标题} （{URL}） {日期}{
摘要}{页面信息}

这里的摘要是搜索引擎生成的一段简短内容，用于描述相应网页的内容；页面信息则在可用时包含额外的轻量级元数据。这种表示方式能让模型在进行任何网页爬取之前，简洁且易读地了解每个候选 URL。

具体而言，对于每个查询，我们都会从 Serper JSON 响应中提取排名前 10 的自然搜索结果。对于每条结果，我们会解析其标题、URL、发布日期（如有）和摘要，并将其序列化为上述 Markdown 格式。随后，我们将格式化后的条目拼接起来，形成最终的搜索结果块，并将其作为网页搜索观测的一部分直接提供给模型。

为提高稳健性，如果发生超时或暂时性 HTTP 错误，每个搜索请求最多会重试五次。为降低 API 成本并提高检索吞吐量，所有搜索结果都会按查询进行缓存；重复查询会复用缓存结果，而不会调用外部搜索 API。

C.5. 网页爬取工具详情

我们实现网页爬取工具时使用开源的 `crawl4AI`[4] 库，该库用于根据给定的 URL 获取并预处理网页内容。该工具旨在保留网页中有用的文本信息，同时减少噪声。这通过一组内容过滤配置来实现：

首先，我们基于 HTML 标签移除非信息性网页组件，包括导航栏、页脚、侧边栏、脚本、样式以及嵌入式媒体元素（例如 nav、footer、script、style、iframe、video）。此步骤消除了与语义推理无关的布局和样板内容。

其次，我们从提取的文本中移除所有超链接和图像。根据经验，我们发现现代大语言模型分词器不太适合处理原始超链接，这些超链接在分词后往往会扩展为异常冗长的令牌序列，并显著增加上下文长度。此外，我们的任务完全基于文本，图像并不能为下游推理提供有用信号。因此，移除链接和图像既能降低令牌开销，也能提高模型输入的稳定性。

最后，清理后的网页内容会被转换为 Markdown 格式。在此过程中，我们还应用了 crawl4AI 提供的额外修剪步骤，移除少于 10 个单词的极短文本片段。这进一步过滤掉了信息量低的内容，并生成紧凑、适合大语言模型处理的网页表示。

D. 案例研究

表 8 展示了 SynPlanResearch-R1 在 GAIA 数据集上的一个案例研究。此示例说明了网页爬取的作用：初始网页搜索结果仅提供高层次的信息，而 SynPlanResearch-R1 会点击进入相关网页，以获取标准化的对面颜色配对，这些信息提供了解决魔方推理问题所需的关键约束。

表 9 和 10 比较了所提出的模型与基线（冷启动强化学习）在同一道 HotpotQA 问题上的表现。SynPlanResearch-R1通过执行更多搜索步骤来收集并核实事实信息，从而得出正确答案；而表10中的基线在收集到充分证据之前就过早作答。

[4]https://github.com/unclecode/crawl4ai

算法 1 SynPlanResearch-R1：计划引导的合成 ReAct 数据合成

需要： 数据集 $\mathcal{D}$，其中每个元素都是一对 (q，gold ans)，由问题 q 及其对应的标准答案 gold ans 组成；展开预算 N；计划长度范围 $[L_{\min}, L_{\max}]$。
确保：冷启动监督微调语料库 $\mathcal{D}_{\text{sft}}$。1：初始化 $\mathcal{D}_{\text{sft}} \leftarrow \emptyset$。2：对于每个（ q，标准答案） $\in \mathcal{D}$ 执行 3：初始化 $\mathcal{S}$（q） $\leftarrow \emptyset$4：// q的候选轨迹 5：对于 $i = 1$ 到 N 执行 6：7：// 第4.1节，第1步：工具计划构建 8：采样 $L \sim$ 均匀分布$\{L_{\min}, \ldots, L_{\max}\}$。9：采样一个工具计划 $P =$（a_1，$\ldots$，a_L），其中包含 $a_1 =$次网络搜索以及 $a_t \sim \mathcal{A}$用于 $t \geq 2$。10：通过将 P注入计划条件模板并追加查询 q来构建初始 ReAct 提示；将其记为 H_0（附录 A.2）。11：12：// 第4.1节，第2步：提示注入的思考 13：对于 $t = 1$ 到 L执行 14：$c_t \leftarrow$ Cue(a_t)，使用特定操作的提示模板（附录 A.3）。15：查询大型推理模型，根据 <think>内部的前缀补全当前 ReAct 步骤：其中 $\Delta\tau_t$ 是思考文本的剩余部分， a_t 是动作（工具调用）16：执行 a_t 以获得观察 o_t （工具响应），17：形成完整的思考块 $\tau_t \leftarrow \langle$think$\rangle c_t \,\|\, \Delta\tau_t$。18：更新历史记录 $H_t \leftarrow H_{t-1} \oplus$（$\tau_t$， a_t， o_t）。19：结束 for 20：让大型推理模型继续执行最后一步（无工具调用），以生成终止的 <think> 和 <answer> 块：并设置 $H_{L+1} \leftarrow H_L \oplus$（$\tau_{L+1}$，ans）。21：从 <answer>标签中解析 ans。22：23：// 第4.1节，第3步：筛选与质量控制 24：运行格式检查(H_{L+1})和答案检查(ans, 标准答案) 25：如果格式检查失败或答案检查失败，则 26：继续 27：结束 if 28：29：// 第4.1节，第4步：思考重写 30：按顺序重写每个思考步骤（附录 A.4）：$\tilde{}H_{L+1} \leftarrow$ REWRITETHOUGHTS(q, H_{L+1})，其中仅重写思考文本，同时保持所有动作、观察和最终答案不变。31：~32：$\mathcal{D}_{\text{sft}} \leftarrow \mathcal{D}_{\text{sft}} \cup H_{L+1}$。33：结束 for 34：结束 for 35：返回 $\mathcal{D}_{\text{sft}}$。

$$(\Delta\tau_t,\ a_t) \sim \mathrm{LRM}\big(\cdot \mid H_{t-1}, \langle \texttt{think} \rangle c_t\big),$$

$$(\tau_{L+1},\ \texttt{ans}) \sim \mathrm{LRM}(\cdot \mid H_L),$$

-

表 8. SynPlanResearch-R1案例研究

问题：一个标准的魔方已被拆成构成其各面的共 26 个小方块。魔方的内部方块已被移除。其中一个被取出的方块只有一个彩色面的中心块、12 个有两个彩色面的棱块，以及 8 个有三个彩色面的角块。所有蓝色小方块都已找到。位于橙色中心块正左、正右、正上和正下方的所有小方块，以及该中心块，都已找到。所有绿色角块都已找到，所有与黄色相邻的绿色小方块也都已找到。对于所有已找到的橙色小方块，其相对面的方块也都已找到。被取出的方块的面上有两种颜色。它们是什么？请使用逗号分隔的列表作答，并按颜色的字母顺序排列。

标准答案：绿色，白色**数据集领域**：GAIA

SynPlanResearch-R1：
<思考> 要解决这个问题，我需要了解魔方小方块的分类（6 个中心块、12 个棱块和 8 个角块），更关键的是标准化的对面颜色配对（例如，蓝色与绿色相对、橙色与红色相对、白色与黄色相对）。我会查找标准的颜色排列。 </思考><工具调用> {”名称”: ”网络搜索”, ”查询”: ”标准魔方的组成” } </工具调用><工具响应> [1] [关于魔方的一切 - Cubelelo] (https://www.cubelelo.com/blogs/cubing/all-about-rubik-s-cube) 结构 - 魔方由一个固定所有部件的球形核心组成，所有部件都相对于中心核心部件移动。除中心块外，共有 20 个部件，其中包括 8 个角块和 12 个棱块。[2] [魔方 - 维基百科] (https://en.wikipedia.org/wiki/Rubik%27s_Cube) 机械结构。一枚标准魔方每边长 5.6 厘米（2+1⁄4 英寸），由 26 个独特的微型方块组成，也称为“小方块”.....

[10] [[PDF]魔方数学]（https://www.fuw.edu.pl/~konieczn/RubikCube.pdf）1992年5月22日 按照标准约定，魔方群将作用于右侧魔方的面集。定义 99。设 G 作用于一个..... </工具响应> <think> 我需要更具体地了解魔方的结构，因为最初的搜索没有提供足够详细的信息，无法说明魔方的不同类型。让我专门搜索魔方中不同类型的块——角块、棱块和中心块——以了解它们的特性以及它们与问题的关系。这将有助于我判断被移除的那块带有两种颜色的方块可能是哪一块。 </think><tool call> {”name”: ”web search”, ”query”: ”魔方的不同类型方块” } </tool call><tool response> [1] [10 不同类型的魔方及其独特特点](https://particula-tech.com/blogs/news/10-different-types-of-rubiks-cubes) 魔方的所有类型[2] [魔方及其变体的图片和名称列表？ : r/Cubers](https://www.reddit.com/r/Cubers/comments/3x5iu1/list_of_rubiks_cubes_and_variations_with_pictures/) 2015年12月17日 我想要一份魔方及其变体的图片列表，并在图片旁标注名称，例如金字塔魔方、3×3、五魔方、魔方复仇、亚历山大之星、空洞魔方等.....[10] [14 史上最酷、最有趣的魔方 - Cubelelo](https://www.cubelelo. com/blogs/cubing/coolest-rubiks-cubes?srsltid=AfmBOoqP TUNA XCo znSnIO An sWqtVxD xNxTMtNz JhJ-D) 4月。面转八面体（FTO）· 2. 镜面魔方 · 3. 五魔方 · 4. 金字塔魔方 · 5. 4×4和5×5（魔方复仇和五阶魔方）· 6. 大型魔方（8×8.....） </tool response think 要解决这个问题，我需要更具体地了解魔方的结构。到目前为止，搜索结果未能提供有关不同类型方块及其颜色配置的详细信息。让我查看维基百科上的魔方词条，其中应该有关于其结构的全面介绍，包括方块的排列方式和各自的颜色。这将帮助我判断被移除的那块带有两种颜色的方块可能是哪一块。</think><tool call> {”name”: ”crawl webpage”, ”url”: ”https://en.wikipedia.org/wiki/Rubik%27s Cube” /tool call tool response ... 自1988年以来，颜色的排列方式已标准化：白色与黄色相对，蓝色与绿色相对，橙色与红色相对 </tool response> .....另外两次网络搜索（为节省空间省略）.....<think> .....被移除的方块是一个带有两种颜色的棱块。根据线索，唯一一个不与橙色中心块相邻且不与其他条件冲突的棱块是**白绿棱块**。按字母顺序排列，颜色为**绿色、白色**。..... </think><answer>绿色，白色 </answer> ✓

表9. SynPlanResearch-R1案例研究**问题**：英国系列动作喜剧电影中，第二部戏仿詹姆斯·邦德特工类型的影片由谁执导？该系列电影的喜剧风格类似于罗温·艾金森饰演的憨豆先生角色。**真实标注**：奥利弗·帕克**数据集领域**：HotpotQA

SynPlanResearch-R1：

<思考> 我需要先找出那部恶搞詹姆斯·邦德类型片的英国动作喜剧系列。问题提到，该系列的喜剧风格类似于罗温·艾金森饰演的憨豆先生，因此我会先搜索这部特定英国系列电影的相关信息。这样我就能确认该系列的“第二部”是哪部电影，然后查明导演是谁。 </思考><工具调用> {”name”: ”网络搜索”, ”query”: ”恶搞詹姆斯·邦德、带有憨豆先生式喜剧风格的英国动作喜剧系列”} </工具调用><工具响应> [1] [你怎么看《憨豆特工》系列？对……的恶搞](https://www.reddit.com/r/JamesBond/comments/1cqk0le/what_are_your_thoughts_on_the_johnny_english/) 2024年5月12日。我可能更喜欢罗温·艾金森饰演的憨豆特工，而不是憨豆先生。第二部电影是我最喜欢的，因为它在幽默和动作之间的平衡最好……[2] [《憨豆特工》（电影系列）- 维基百科] (https://en.wikipedia.org/wiki/Johnny_English_(film_series)) 《憨豆特工》是一部恶搞詹姆斯·邦德特工类型的间谍动作喜剧电影系列，由罗温·艾金森饰演主角……[10] [如果詹姆斯·邦德遇到某人会作何反应……- Quora] (https://www.facebook.com/.../1271602630992141) 憨豆先生是个喜剧演员，和邦德在克雷格时代执行的严肃任务毫无关系。《憨豆特工》系列算是对……的恶搞 </工具响应><思考> 根据搜索结果，我已确认，这部带有憨豆先生式喜剧风格、恶搞詹姆斯·邦德的英国动作喜剧系列是《憨豆特工》。搜索结果提到，该系列是一部恶搞詹姆斯·邦德特工类型的间谍动作喜剧电影系列，由同样饰演过憨豆先生的罗温·艾金森主演。现在我需要找出该系列第二部电影的导演是谁。我来专门搜索《憨豆特工》第二部电影的导演信息。 </思考><工具调用> {”name”: ”网络搜索”, ”query”: ”《憨豆特工》第二部电影的导演”}</工具调用><工具响应> [1] [《憨豆特工2》- 维基百科](https://en.wikipedia.org/wiki/Johnny_English_Reborn) 《憨豆特工2》是一部2011年英国间谍动作喜剧电影，由奥利弗·帕克执导，哈米什·麦科尔编剧，故事由William Davies创作。[2] [《憨豆特工2》（2011）- IMDb](https://www.imdb.com/title/tt1634122/) IMDb 条目：《憨豆特工2》。影片由奥利弗·帕克执导，罗温·艾金森主演，讲述憨豆特工面对国际威胁的间谍恶搞故事……

[10] [《憨豆特工2》（2011年）导演：奥利弗·帕克主演](https://www.facebook.com/.../1271602630992141) 2025年6月25日。一则 Facebook 帖子总结了《憨豆特工2》（2011年），确认奥利弗·帕克担任导演，并列出了包括罗温·艾金森、Gillian Anderson、Rosamund Pike 和 Dominic West 在内的主要演员。 </工具响应>

<思考> 根据搜索结果，我确认了这部模仿詹姆斯·邦德、带有憨豆先生式喜剧风格的英国动作喜剧系列是《憨豆特工》。我还发现，该系列有多部电影，第二部由奥利弗·帕克执导。不过，我应该核实这些信息，并进一步了解这个系列及其电影。让我搜索《憨豆特工》电影系列的更具体信息，以确认第二部电影的导演。 </思考>

<工具调用> {”名称”: ”网络搜索”, ”查询”: ”《憨豆特工》电影系列导演”} </工具调用><工具响应> [1] [《憨豆特工》- 维基百科]（https://en.wikipedia.org/wiki/Johnny_English）《憨豆特工》是一部于 2003 年上映的英国间谍动作喜剧电影，由 Peter Howitt 执导，Neal Purvis、Robert Wade 和 William Davies 编剧。

[2] [《憨豆特工》电影系列 - 维基百科](https://en.wikipedia.org/wiki/Johnny_English_(film_series)) 《憨豆特工》电影系列概述。该系列恶搞詹姆斯·邦德特工类型，并由多位导演执导，包括 Peter Howitt 、奥利弗 · 帕克和 David Kerr。...[10] [《憨豆特工》系列 — 电影数据库（ TMDB）](https://www.themoviedb.org/collection/86117-johnny-english-collection?language=en-US) TMDB 系列页面介绍了《憨豆特工》电影，将其描述为恶搞詹姆斯·邦德类型的英国动作喜剧，由罗温·艾金森饰演主角。 </工具响应> ... 还有一次网络搜索（为节省空间省略）...<思考> 这部英国动作喜剧间谍恶搞系列指的是由罗温·艾金森主演的《憨豆特工》电影。该系列的第二部电影是《憨豆特工2》（2011年），由奥利弗·帕克执导。... </思考> <答案>奥利弗·帕克 </答案> ✓

表 10。 采用朴素冷启动监督微调的强化学习案例研究

问题：英国系列动作喜剧电影中，哪位导演执导了第二部戏仿詹姆斯·邦德特工类型、且喜剧风格类似于罗温·艾金森饰演的憨豆先生这一角色的电影?

真实标注：奥利弗·帕克

数据集领域：HotpotQA

SynPlanResearch-R1：

<思考> 好，我们来想想，哪位导演执导了那部恶搞詹姆斯·邦德的英国动作喜剧电影的第二部。用户提到它类似于罗温·艾金森主演的《憨豆先生》，所以我得回想或查找这个系列的名字。首先，我记得有一部英国电影叫《急速秀》，但我不确定它是否就是那个。等等，也许是《裸体炸弹》？不，那是另一种类型。哦，对了！有一部电影叫《憨豆特工》，是一部恶搞詹姆斯·邦德的喜剧。让我确认一下。《憨豆特工》于1999年上映，续集是2000年的《憨豆特工：回归》..... </think>

<工具调用> {"name": "网络搜索", "query": "英国 詹姆斯·邦德 恶搞 动作 喜剧 系列"} </tool call>

<工具响应> [1] [詹姆斯·邦德恶搞作品及衍生作品列表](https://en.wikipedia.org/wiki/List_of_James_Bond_parodies_and_spin-offs) 维基百科的列表，汇总了恶搞詹姆斯·邦德系列或由其衍生的电影和电视作品，其中包括《六月的酷暑》（1964年）和《继续间谍》（1964年）等英国间谍喜剧。

[2] [恶搞与模仿 — 詹姆斯·邦德维基 - Fandom]

(https://jamesbond.fandom.com/wiki/Parodies_%26_模仿）詹姆斯·邦德恶搞和模仿作品概览，重点介绍一些知名例子，例如《王牌大贱谍》电影系列及其对邦德角色和套路的戏仿。

:

[10] [间谍喜剧电影](https://www.imdb.com/list/ls040234091/)一个IMDb间谍喜剧电影列表，其中包括由罗温·艾金森主演的《憨豆特工》系列电影，例如《憨豆特工》（2003年）、《憨豆特工2》（2011年）和《憨豆特工3》（2018年）。</工具响应>

<思考> 好的，用户询问的是一部英国系列电影中第二部作品的导演。该系列以类似罗温·艾金森饰演的憨豆先生的喜剧风格恶搞詹姆斯·邦德类型。从网络搜索结果来看，《憨豆特工》系列被提及为一部恶搞詹姆斯·邦德的间谍动作喜剧，罗温·艾金森是主角。这一定就是用户所指的系列。现在，第一个结果提到了《詹姆斯·邦德恶搞作品列表》，第四个结果是维基百科的《憨豆特工》（电影系列）页面。让我查看一下。维基百科页面称，《憨豆特工》有两部主要电影：第一部于1999年上映，第二部《憨豆特工：回归》于2000年上映。接下来，我需要查明第二部电影的导演。最初的搜索没有提到导演，因此我应该再次进行网络搜索，专门查找《憨豆特工：回归》的导演。让我再次使用网页搜索工具进行查询。</思考>

<工具调用> {"名称": "网络搜索", "查询": "《憨豆特工 : 回归》"} </工具调用>

<工具响应> [1] [《憨豆特工3》（2018年）- IMDb](https://www.imdb.com/title/tt6921996/) IMDb 页面：《憨豆特工*3*》（2018年）。该片由 David Kerr 执导，罗温·艾金森主演，延续了英国间谍动作喜剧恶搞系列。

[2] [《憨豆特工3》（2018年）- 完整演职员表 - IMDb]

(https://www.imdb.com/title/tt6921996/fullcredits/) 该片的完整演职员表，确认 David Kerr 担任导演、William Davies 担任编剧，并提供详细的制作信息。。

:

[10] [David Kerr on Rowan 阿特金森, 约翰尼 英国 再度 出击, and 喜剧] (https://weliveentertainment.com/welivefilm/大卫·克尔谈罗温·艾金森、《憨豆特工》与喜剧/) 一篇采访 ar- 发表于2018年11月6日的文章，导演 David Kerr 在文中谈及他参与执导的《憨豆特工*2*》: 再次，他与罗温·艾金森的合作经历，以及他在喜剧领域的背景。 </工具响应>

<思考> 好的，用户问的是恶搞詹姆斯·邦德系列的英国系列电影第二部的导演，特别提到了罗温·艾金森式的憨豆先生风格喜剧。这一类型，具体来说是罗温·艾金森的憨豆先生式喜剧。根据此前的搜索后，我们确定所说的系列是《憨豆特工》。首次搜索确认，《憨豆特工 : 回归》是该系列的第二部电影。查看当前关于导演的搜索结果，《憨豆特工：回归》的多处来源都提到 David Kerr 是导演。例如，IMDb 条目将 David Kerr 列为导演，其他评论和文章也证实了这一点。第二部电影《回归》于 2000年 上映，而各个平台列出的导演始终是 David Kerr。我应该核实是否存在相互矛盾的信息，但这里所有来源的说法一致。因此，答案应该是 David Kerr。 </思考>

<答案>David Kerr </答案> ×