WebDancer: Towards Autonomous Information Seeking Agency

如何拥有像DEEP RESEARCH那样的信息检索能力?

代码：HTTPS://GITHUB.COM/ALIBABA-NLP/WEBAGENT

Jialong Wu*, Baixuan Li*, Runnan Fang*, Wenbiao Yin*†, Liwen Zhang,
Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang†, Pengjun Xie, Fei Huang, Jingren Zhou

Tongyi Lab , Alibaba Group

Correspondence to: wujialongml@gmail.com

{yinwenbiao.ywb,yongjiang.jy}@alibaba-inc.com

简介

WebDancer 通过 ReAct-style 推理轨迹建模、可控复杂度的 TIR 数据生成机制，以及 SFT + DAPO 的两阶段训练策略，实现了面向信息检索任务的多工具集成推理。

1. ReAct范式的推理和工具交互流程：即以交替的 { <think> 思考、<tool> 调用 和 <observe> 工具响应}三元组的形式组织推理轨迹。
2. 覆盖广度和可控难度的TIR推理数据生成机制：a) CRAWLQA通过爬取高质量网页内容，使用GPT-4o合成 COUNT、MULTI-HOP、INTERSECTION 等类问题，提升数据的知识覆盖与任务广度；b) E2HQA 采用逐步“反向构造”策略，利用搜索和重写操作逐步构造出更复杂的多跳问题，可控地增加数据推理难度。
3. 两阶段训练流程：SFT + DAPO

背景

尽管LLM在问答与推理任务中取得了显著进展，但在开放领域的信息获取场景中仍面临知识截止与幻觉问题。当前主流解决方案是将搜索工具引入推理流程，形成Tool-Integrated Reasoning（TIR）框架，使模型在生成过程中可动态调用外部工具以弥补知识盲区。

如果想进一步具备类似Deep Research式的深度信息检索与推理能力，该如何做呢？WebDancer从可控复杂度的(query, answer)构造、推理轨迹生成和两阶段训练流程，进行了探索。

实验设置

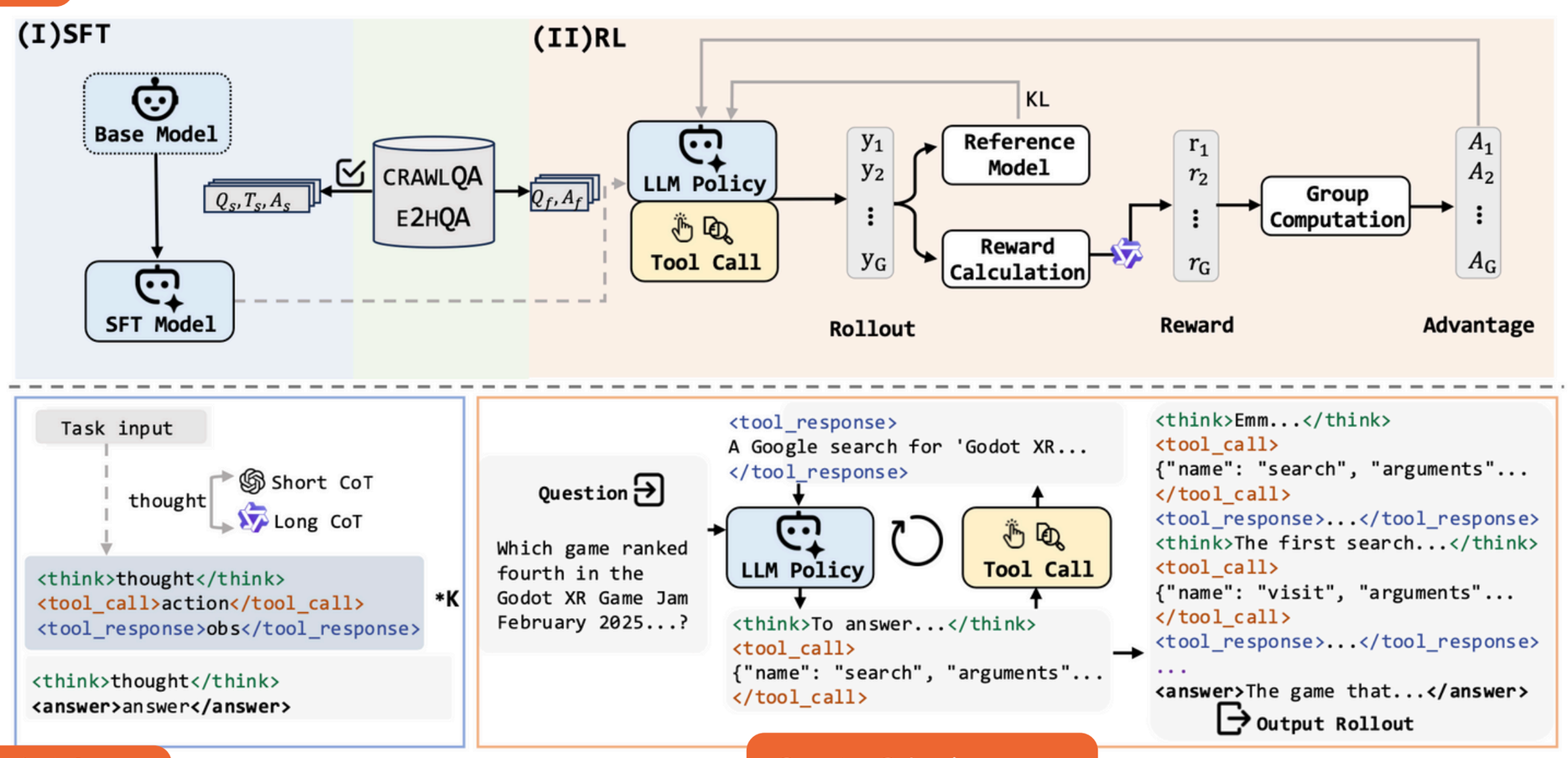
信息检索任务: GAIA和WEBWALKERQA

两个TOOL：SEARCH 和访问网页

- 实验对象：Qwen2.5 7B/32B, QwQ 32B
- 强化学习算法：DAPO
- ORM Reward设计: format reward和answer reward两项。两个reward都是二值型(取值0或1)，用LLM-as-Judge计算answer reward而非计算EM或F1

$$R(\hat{y}_i, y) = 0.1 * score_{\text{format}} + 0.9 * score_{\text{answer}}$$

训练框架



部分实验结果

Close-Sourced Agentic Frameworks									
	OpenAI DR	74.3	69.1	47.6	67.4	-	-	-	-
Open-sourced Agentic Frameworks									
Qwen-2.5-7B	Search-o1	23.1	17.3	0.0	17.5	-	-	-	-
	R1-Searcher	28.2	19.2	8.3	20.4	-	-	-	-
Qwen-2.5-32B	Search-o1	33.3	25.0	0.0	28.2	-	-	-	-
QwQ-32B	Search-o1	53.8	34.6	16.7	39.8	43.1	35.0	27.1	34.1
	WebThinker-Base♣	51.2	43.4	8.3	41.7	47.5	33.2	25.0	33.6
	WebThinker-RL♣	53.8	44.2	8.3	43.7	46.2	39.2	28.7	37.2
ReAct Agentic Frameworks									
Qwen-2.5-7B	Vanilla ReAct	28.2	15.3	0.0	18.4	28.1	31.2	16.0	24.2
	WebDancer‡	41.0	30.7	0.0	31.0	40.6	44.1	28.2	36.0
Qwen-2.5-32B	Vanilla ReAct	46.1	26.9	0.0	31.0	35.6	38.7	22.5	31.9
	WebDancer‡	46.1	44.2	8.3	40.7	44.3	46.7	29.2	38.4
QwQ-32B	Vanilla ReAct	48.7	34.6	16.6	37.8	35.6	29.1	13.2	24.1
	WebDancer†	**56.4**	**48.1**	**25.0**	**46.6**	**49.4**	**55.0**	**29.6**	**43.2**
GPT-4o	Vanilla ReAct	51.2	34.6	8.3	34.6	34.6	42.0	23.9	33.8

CONCLUSION

Deep Research展现出的深度信息检索与融合能力让人眼馋，本文从训练数据构造与多工具TIR的角度出发，提出了自己的探索方案，如果再加上一个自动生成报告的模块，那就更完善啦。不过有一点我没想明白：作者将TIR推理轨迹表示为ReAct 格式，即（think, act, act response）组成的三元组序列。事实上，很多TIR工作本来就采用类似的结构化标签（如 <think>、<query>、<doc> 等）来区分推理过程中的不同内容，形式上也符合ReAct的定义。所以，ReAct在这里是一种写作封装还是确实有所不同呢？不管怎样，我相信Deep Research方向会持续涌现出更多优秀的工作，令人期待。