ReSearch: Learning to *Reason* with *Search* for LLMs via Reinforcement Learning

在RLVR背景下，如何训练LLM做SEARCH-AND-REASONING

Mingyang Chen[1], **Tianpeng Li**[1], **Haoze Sun**[1], **Yijie Zhou**[1], **Chenzheng Zhu**[1], **Haofen Wang**[2], **Jeff Z. Pan**[3], **Wen Zhang**[4], **Huajun Chen**[4], **Fan Yang**[1]*, **Zenan Zhou**[1], **Weipeng Chen**[1]

[1]Baichuan Inc. [2]Tongji University [3]The University of Edinburgh [4]Zhejiang University

{chenmingyang, yangfan}@baichuan-inc.com

https://github.com/Agent-RL/ReSearch

简介

本文提出了ReSearch (Reason with Search)，探索在多跳QA任务上，延续RLVR做法，让llm学会使用search tool提升自己的能力。

背景

本文比之前读过的R1-Searcher和Search-R1要晚几天，背景知识就直接复制：

LLM与搜索引擎（search engine）结合可以扩展其内部知识，如何结合呢？一种方法是RAG，通过搜索引擎的检索结果来扩展prompt；另一种是把搜索引擎看作一种tool，让LLM学会使用search tool。

让LLM使用tool，最简单的方法是写prompt template，比如解释下search tool可以做什么，再举几个使用tool的prompt的例子，类似CoT。还可以对LLM做fine-tuning，训练它学会使用search tool，本文聚焦用RL做tuning，既让LLM提升推理能力又学会使用search tool

实验设置

多跳QA任务

BASE和INSTRUCT模型用不同的PROMPT TEMPLATE
引导输出指定格式的RESPONSE

- 实验对象：Qwen2.5-7B/32B Base/Instruct
- 强化学习算法：GRPO　FLASHRAG作为SEARCH TOOL
- ORM格式的RLVR reward function：答案是否正确（F1）+format reward。包含kl loss项

$$r = \begin{cases} \mathrm{f1}(a_{\text{pred}}, a_{\text{gt}}), & \text{if f1 score is not 0} \\ 0.1, & \text{if f1 score is 0 and format is correct} \\ 0, & \text{if f1 score is 0 and format is incorrect} \end{cases}$$

训练流程

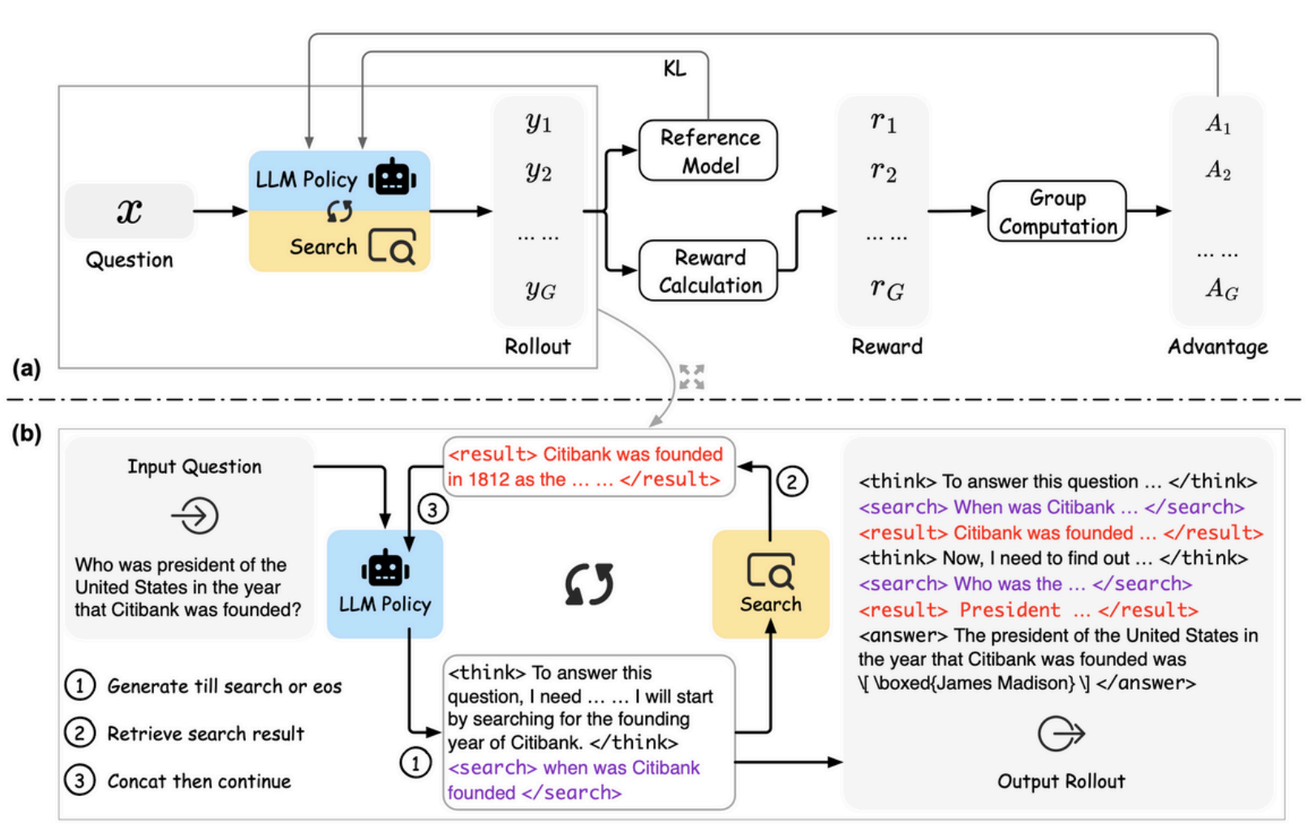

Figure 2: The training overview of *ReSearch*. (a) The GRPO pipeline. (b) The details of the rollout generation process.

检索结果要MASK，不参与LLM参数更新

部分实验结果

Model	HotpotQA		2Wiki		MuSiQue		Bamboogle	
	EM	*LJ*	*EM*	*LJ*	*EM*	*LJ*	*EM*	*LJ*
Qwen2.5-7B(-Instruct)								
Naive Generation	19.18	30.64	25.76	27.87	3.76	10.38	10.40	22.40
Naive RAG	31.90	49.59	25.78	29.52	6.21	12.78	20.80	32.00
Iter-RetGen	34.36	52.22	27.92	31.86	8.69	16.14	21.60	35.20
IRCoT	30.33	52.06	21.57	30.65	6.99	14.19	24.80	36.80
ReSearch-Qwen-7B	40.57	60.26	44.67	50.06	21.68	32.19	**43.20**	**54.40**
ReSearch-Qwen-7B-Instruct	**43.52**	**63.62**	**47.59**	**54.22**	**22.30**	**33.43**	42.40	**54.40**
Qwen2.5-32B(-Instruct)								
Naive Generation	24.63	38.26	27.23	29.68	6.12	14.23	18.40	29.60
Naive RAG	36.46	55.73	30.38	34.87	9.27	15.97	23.20	40.80
Iter-RetGen	39.81	58.80	33.64	38.22	12.49	20.11	29.60	44.80
IRCoT	28.44	55.44	13.53	29.50	7.82	18.20	31.20	47.20
ReSearch-Qwen-32B	42.77	64.27	38.52	45.59	**26.40**	37.57	54.40	66.40
ReSearch-Qwen-32B-Instruct	**46.73**	**67.70**	**44.90**	**50.30**	**26.40**	**38.56**	**56.80**	**67.20**

思考

本文比R1-SEARCHER和SEARCH-R1晚几天发布，类似的内容已经思考过了，暂时没有什么额外要思考的