R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning

在RLVR背景下，用两阶段法训练LLM做SEARCH-AND-REASONING

Huatong Song[1]*, **Jinhao Jiang**[1]*, **Yingqian Min**[1], **Jie Chen**[1], **Zhipeng Chen**[1], **Wayne Xin Zhao**[1]†, **Lei Fang**[2], **Ji-Rong Wen**[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China.
[2]DataCanvas Alaya NeW
{songhuatong123, jiangjinhao}@ruc.edu.cn
batmanfly@gmail.com

简介

开源代码：HTTPS://GITHUB.COM/RUCAIBOX/R1-SEARCHER

本文提出了R1-Searcher，探索在多跳问答场景，使用RLVR的做法，让llm学会使用外部检索系统/搜索引擎tool。作者设计了两阶段的训练流程，有点课程学习的感觉，第一阶段主要鼓励llm在reasoning时多多调用外部检索系统，第二阶段才让llm真正去得到正确答案。为了匹配两阶段训练，训练集样本也根据难度划分，基本上第二阶段的样本要更难推理。

背景

本文比之前读过的Search-R1要早几天，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

训练集：两个多跳问答数据集HOTPOTQA和2WIKIMULTIHOPQA

检索结果要MASK，不参与LLM训练

- 实验对象：Qwen-2.5-7B-Base 和 Llama-3.1-8B-Instruct
- 强化学习算法：GRPO/Reinforce++
- ORM形式的RLVR reward function：第一阶段不管answer是否正确，包含format reward和推理阶段是否调用了search tool两项

$$R_{retrieval} = \begin{cases} 0.5, & n \geq 1 \\ 0, & n = 0 \end{cases} \qquad R_{format} = \begin{cases} 0.5, & \text{if the format is correct} \\ 0, & \text{if the format is incorrect} \end{cases}$$

第二阶段包括format reward和answer是否正确两项。后者用F1指标

$$R_{answer} = \frac{2 * IN}{PN + RN} \qquad R'_{format} = \begin{cases} 0, & \text{if the format is correct} \\ -2, & \text{if the format is incorrect} \end{cases}$$

RESPONSE格式

System Prompt for Base Model

The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the final answer. The output format of reasoning process and final answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., "<think> reasoning process here </think><answer> final answer here </answer>". During the thinking process, **the Assistant can perform searching** for uncertain knowledge if necessary with the format of "<|begin_of_query|> search query (only list keywords, such as "keyword_1 keyword_2 ...")<|end_of_query|>". **A query must involve only a single triple**. Then, the search system will provide the Assistant with the retrieval information with the format of "<|begin_of_documents|> ...search results... <|end_of_documents|>".

部分实验结果

Zero-Shot		Naive Generation	0.208	0.268	0.326	0.254	0.144	0.168	0.068	0.096
		Standard RAG	0.334	0.398	0.336	0.212	0.168	0.216	0.104	0.098
Branching		SuRe	0.266	0.346	0.122	0.262	0.160	0.192	0.106	0.144
		REPLUG	0.290	0.348	0.334	0.204	0.168	0.232	0.078	0.090
Summary		LongLLMLingua	0.314	0.382	0.304	0.294	0.168	0.216	0.088	0.100
		RECOMP	0.318	0.380	0.324	0.322	0.104	0.160	0.112	0.126
Llama		Selective-Context	0.296	0.358	0.266	0.204	0.144	0.200	0.092	0.104
	Adaptive	SKR	0.300	0.372	0.336	0.212	0.176	0.208	0.100	0.112
	RAG-CoT	Self-Ask	0.316	0.408	0.306	0.322	0.360	0.432	0.222	0.226
		Iter-RetGen	0.302	0.362	0.310	0.224	0.144	0.176	0.084	0.084
		IRCoT	0.210	0.146	0.338	0.312	0.120	0.104	0.060	0.042
	Test-Time	CR-Planer	0.332	0.350	0.420	0.350	0.304	0.336	0.144	0.098
		ReARTeR	0.424	0.434	0.470	0.364	0.438	0.484	0.244	0.252
	Reasoning	Marco-o1	0.352	0.348	0.442	0.184	0.224	0.200	0.134	0.104
		Skywork-o1	0.306	0.256	0.344	0.190	0.176	0.160	0.092	0.060
Llama	**RL**	R1-Searcher	0.648	0.746	0.594	0.628	0.504	**0.544**	0.254	0.282
Qwen	**RL-Zero**		**0.654**	**0.750**	**0.636**	**0.650**	**0.528**	**0.544**	0.282	**0.314**

Method	HotpotQA			2Wiki			Bamboogle			Avg (CEM)
	EM	CEM	F1	EM	CEM	F1	EM	CEM	F1	
GRPO	53.0	60.5	68.6	58.0	60.5	63.0	48.0	56.0	60.5	59.0
Reinforce++	58.4	64.8	70.6	57.5	61.5	62.9	44.0	50.4	57.1	58.9

Table 3: Performance comparison of Llama-3.1-8B-Instruct trained using GRPO and Reinforce++ on three multi-hop QA benchmarks.

思考

Stage	Dataset	Easy	Medium	Difficult
Stage-1	HotpotQA	-	200	-
	2WikiMultiHopQA	-	150	-
Stage-2	HotpotQA	-	2561	2000
	2WikiMultiHopQA	-	1087	2500

Table 1: The information of the data used during RL training.

本文要比SEARCH-R1早几天发布，做的事情类似，因此同样的思考内容就不写了。只说下两阶段训练，大部分推理+TOOL USING的工作都是一个阶段的训练流程。本文第一个阶段侧重让LLM按照给定的FORMAT输出（不同内容用<某种TAG>标记）以及鼓励LLM在推理时多多给出QUERY让外部系统去调用SEARCH TOOL。第二阶段才计算ANSWER是否正确。从训练数据来看，第一阶段用到的训练集挺小的，数据难度集中在中等。至于是否有必要细化成两个阶段或者多个阶段训练，估计还是要实践为主，就像LLM预训练/后训练到底分多少个阶段，每个团队都有自己的考量。