arxiv.org/abs/2507.21836

AutoTIR: Autonomous Tools Integrated Reasoning via Reinforcement Learning

AutoTIR：通过三阶段RL训练具备多工具自主选择的能力

Yifan Wei[1,2], Xiaoyan Yu[3], Yixuan Weng[4], Tengfei Pan[2], Angsheng Li[1†], Li Du[2†]

[1]Beihang University [2]BAAI [3]Beijing Institute of Technology [4]Westlake University
weiyifan@buaa.edu.cn, angsheng@buaa.edu.cn, duli@baai.ac.cn

代码：https://github.com/weiyifan1023/AutoTIR

简介

本文提出AutoTIR，即用RLVR范式训练multi-tool TIR，作者的目标是让lrm学会自主选择工具的同时，尽量保留其原有的指令追随能力。我个人认为本文的训练流程最能体现出作者的意图，将GRPO训练分为三个阶段，每个阶段对应一种工具明确的数据集：1) MuSiQue数据集，强化lrm在知识密集型任务中调用Wiki search工具；2)ToRL和Math-DAPO数据集，引导lrm在数学推理任务中使用Python解释器；3) NQ和指令跟随数据，训练lrm识别哪些问题可以直接用语言能力解决，避免滥用工具。
通过这种按任务类型分阶段训练的方式，让lrm逐步具备“用或不用工具、用哪种工具”的能力。

背景

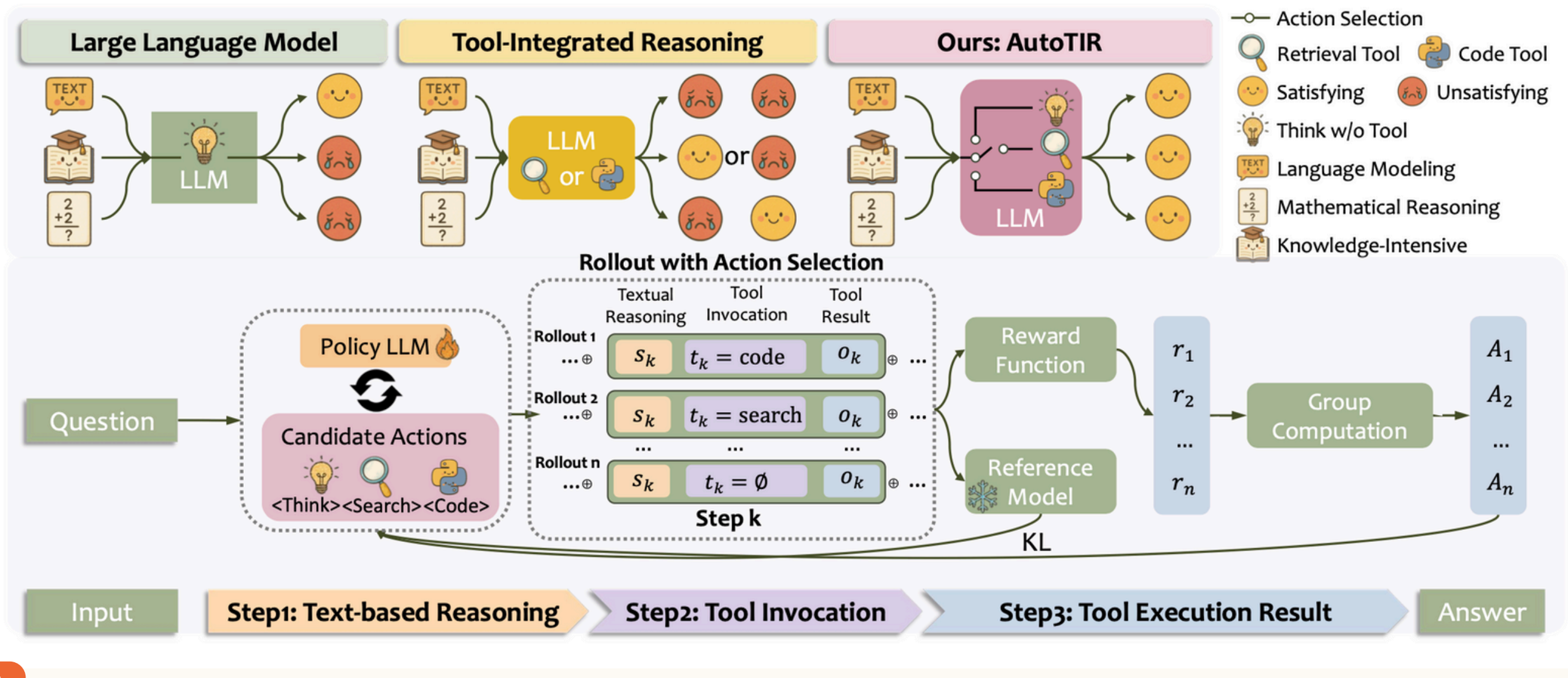
本文属于TIR(tool-integrated reasoning) with RLVR领域的工作，更确切地说是multi-tool TIR，因为要让lrm学会Wikipedia search和Python解释器两个工具，但需要注意的是，训练数据中的每个query最多只需要使用一种工具，因此虽然经过GRPO训练后的lrm具备使用两种工具的能力，但在推理过程中，对于单个query，它的reasoning trajectory 中仍然只会调用一种工具(或完全不调用工具)，并不会出现两种工具组合调用。

实验设置

- 两个工具：Wikipedia search和Python解释器
- 三类训练集：学会调用Wiki search的MuSiQue、学会调用Python解释器的ToRL和Math-DAPO、保持语言模型能力防止滥用工具的NQ和指令跟随数据集
- RL算法：GRPO reward: $r = 0.1 \times r_{\text{act}} + 0.9 \times r_{\text{out}}$

Model	Knowledge-Intensive Domain				Mathematical Domain				Open Domain		AVG
	HotpotQA EM	2Wiki EM	MuSiQ EM	Bamb EM	AIME24 EM	AIME25 EM	MATH500 EM	GSM8K EM	LogiQA Acc	IFEval SAcc	
Qwen2.5-7B-Instruct	19.27	25.49	3.60	10.40	0.00	0.00	20.40	18.57	52.99	67.65	21.84
Naive RAG	32.18	25.62	6.41	19.20	0.00	0.00	17.40	17.13	48.54	71.35	23.78
Iter-RetGen	34.65	27.81	8.23	20.00	3.33	0.00	17.18	16.83	48.69	**71.53**	24.83
IRCoT	30.52	21.29	7.16	22.40	0.00	0.00	11.80	31.46	35.79	28.84	18.93
SimpleRL-Zero	4.42	12.03	1.37	14.40	26.67	**16.67**	60.00	84.76	35.48	7.76	26.36
Eurus-2-7B-PRIME	11.52	22.53	1.82	12.00	16.67	13.33	62.00	**90.07**	43.78	20.52	29.42
ToRL	1.12	0.49	0.37	4.00	**33.30**	10.00	58.40	81.96	39.02	13.12	24.18
Search-R1	35.41	31.23	15.18	40.00	13.33	3.33	36.00	56.18	47.31	14.60	29.26
IKEA	26.75	23.51	14.23	23.20	13.33	3.33	42.40	48.14	50.23	28.65	27.38
ReSearch	42.17	**44.79**	21.27	41.60	0.00	0.00	32.00	47.54	37.94	19.22	28.65
AutoTIR	**43.15**	44.47	**23.58**	**43.20**	**33.33**	**16.67**	**62.60**	88.48	**53.56**	51.02	**46.01**

AutoTIR框架



思考

说实话读完论文后，我一开始并没有看出本文与其他TIR工作在“创新”上的区别，花了不少时间思考AutoTIR中的“Auto”究竟体现在哪，直到在实验章节才逐渐理清思路：作者通过精心设计的三阶段训练流程，分别用不同类型的数据教会模型在特定任务中调用搜索工具、代码工具，以及不调用工具推理回答。使得训练后的lrm能够根据query自主判断是否需要工具、以及该调用哪一种工具，体现出了Autonomous