arxiv.org/abs/2507.18973

A Toolbox, Not a Hammer — Multi-TAG: Scaling Math Reasoning with Multi-Tool Aggregation

Multi-TAG: 多工具ensemble的prompt method for TIR

Bohan Yao[1,2] **Vikas Yadav**[1]
[1]ServiceNow AI [2]University of Washington
s1104@cs.washington.edu

计划开源

简介

本文提出Multi-TAG(Multi-Tool Aggregation)框架，用于数学领域的TIR(Tool-Integrated Reasoning)。简单来说，Multi-TAG属于prompt-based method for TIR, 不需要对llm进行tuning (sfr/rlvr)，与之前TIR工作的最大区别是，以往每个reasoning step至多调用一个tool，而Multi-TAG要求一个step并行的调用多个tool，然后通过“答案一致性 + 最短推理”两阶段机制筛选出最合适的那个reasoning step，然后继续做下一个step的inference。
注意本文的multi-tool是多个同类型的tool做ensemble learning而不是像之前看过的multi tool之间配合(比如搜索引擎和检索tool)

背景

本文属于TIR(Tool-Integrated Reasoning)范畴的工作，TIR的背景就不多说了。

实验设置

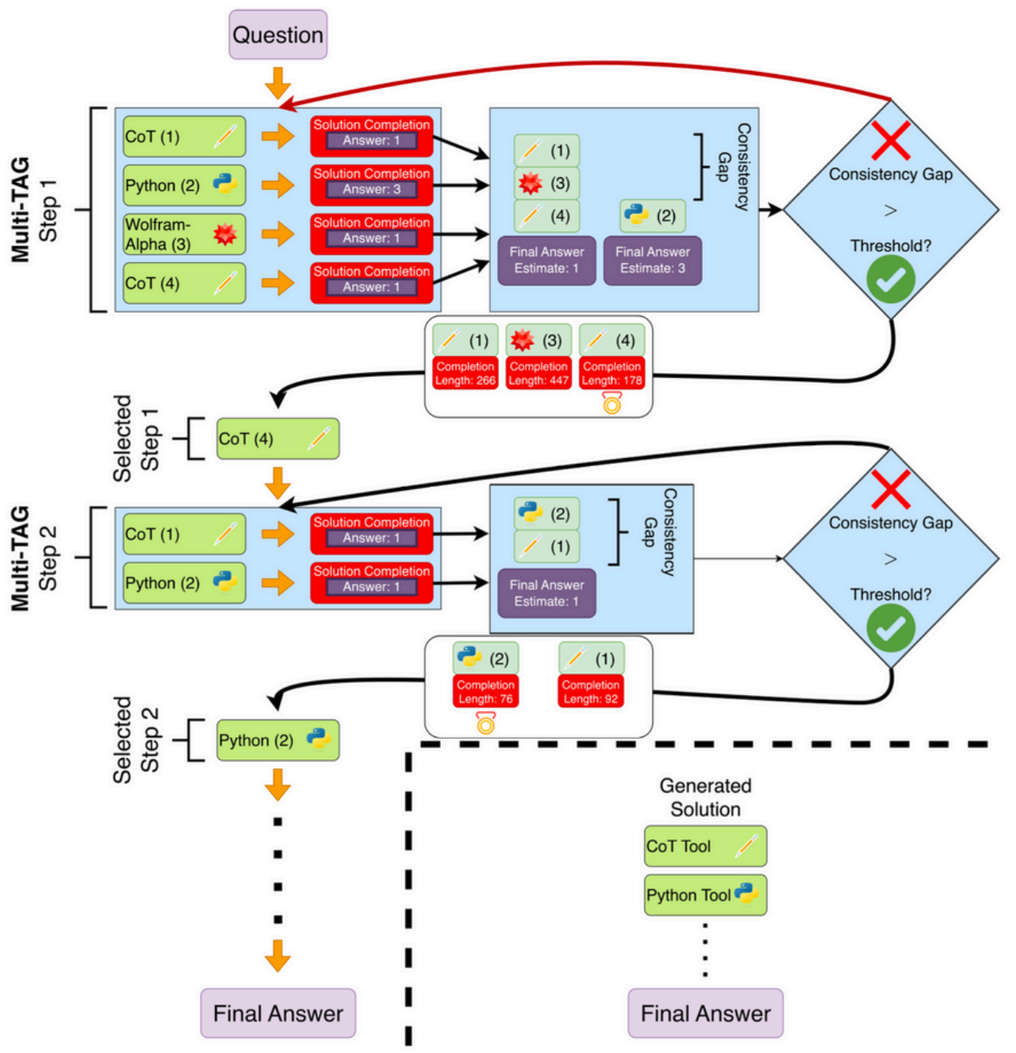
- 数学推理任务；实验对象：GPT-4o、Llama-3.3-70B、Llama-3-70B
- 三个tool：基于纯语言的CoT reasoning、Python 解释器、WolframAlpha

GPT-4o	CoT	79.6%	10.0%	47.0%	32.5%	42.3%
	Python	66.2%	22.2%	50.6%	30.2%	42.3%
	WolframAlpha Query	54.4%	4.4%	22.9%	16.6%	24.6%
	CoT MV	81.8%	12.2%	49.4%	36.5%	45.0%
	Python MV	74.2%	28.9%	59.0%	34.8%	49.2%
	WolframAlpha MV	56.2%	5.6%	22.9%	16.9%	25.4%
	CoT + Python + WolframAlpha MV	86.0%	22.2%	60.2%	38.2%	51.7%
	PAL	64.6%	20.0%	44.6%	28.8%	39.5%
	PoT	51.2%	15.6%	36.1%	19.3%	30.6%
	ToRA	73.0%	17.8%	42.2%	32.1%	41.3%
	MATHSENSEI	73.4%	5.6%	43.4%	28.9%	37.8%
	ReAct	75.2%	28.9%	45.8%	32.1%	45.5%
	Multi-TAG (Ours)	**87.0%**	**34.4%**	**71.1%**	**44.1%**	**59.2%**

Multi-TAG

以第i个reasoning step为例，Multi-TAG 会按照如下流程进行：

- 初始化阶段：原始问题p和S_{i-1} = [s_1, s_2, ..., s_{i-1}]
- 并行调用多个tool，每个tool生成一个step，然后有一个llm基于p、S_{i-1}和tool生成的step进行续写后面的step得到一个answer
- 早停机制：Multi-TAG会统计目前为止的所有answer，计算其consistency gap，一旦gap值超过设置的阈值，说明llm对推理出来的answer已经很有信心了， 可以停止当前step的流程了
- 当前step筛选：先统计所有answer的频次，将频率最高的answer的对应的step都保留，再将其中token数量最少(奥卡姆剃刀)的step作为s_i



思考

读完之后最直接的感受是，inference成本增加会很大啊，可以类比单个model vs ensemble，总之如何取舍效果和效率，看情况而定吧。

@机器爱学习