arxiv.org/abs/2508.12903

A Stitch in Time Saves Nine: Proactive Self-Refinement for Language Models

Jinyi Han[♡], **Xinyi Wang** [♢], **Haiquan Zhao**[♠], **Tingyun li**[♢], **Zishang Jiang** [♢], **Sihang Jiang** [♠], **Jiaqing Liang** [♢], **Xin Lin** [♡], **Weikang Zhou** [♣], **Zeye Sun** [♣], **Fei Yu** [♣], **Yanghua Xiao**[♠] *,
[♡]Shanghai Institute of Artificial Intelligence for Education, East China Normal University
[♢]School of Data Science, Fudan University
[♠]College of Computer Science and Artificial Intelligence, Fudan University
[♣] Antgroup
jinyihan099@gmail.com, xinywang24@m.fudan.edu.cn

简介

代码地址: github.com/JinyiHan99/Proactive-Self-Refine-in-LLMs/

本文提出ProActive Self-Refinement (PASR)训练方法，可以简单理解为RLVR，训练后的llm能在response generation 的过程中主动判断是否需要修正(whether)、何时进行修正(when)以及如何修正(how)。关键问题是如何设计reward function？本文reward分为三部分：1) format reward，约束response必须满足预定义的结构要求，具体可以看论文的system prompt，重点是在<think>内部增加了<refine> tag ; 2) acc reward, 由于训练数据来自开放域QA，因此用judge llm计算acc；3) 关键的refinement reward，作者的思路是针对query同时采样多个不带refine的rollout计算平均acc，然后和带有refine的acc去比较，只有在修正带来acc提升时才给出奖励，否则惩罚，防止无意义的refine。

背景

本文属于llm self-refinement方向的工作，现有的self-refinement方法基本是post-hoc(事后修正)类型，比如常用的critic-refine范式：llm先生成一个初始的response，再让critic去得到可改进的建议，llm根据建议修正response，以上经历多轮迭代得到最终的response。
这类方法存在局限，首先critic的修正相对比response的生成是滞后的，其次到底要修正几轮不明确，一般设置轮数超参数，比较低效。本提出了一种让llm在generation过程中主动refine的思路。

实验设置

- 训练集：来自alpaca_evol_instruct_70k的4W条(question, answer)数据
- 评测集：MMLU、DROP、GSM8K、MATH、AIME24等
- 模型：Qwen2.5-7B和Qwen3-8B

$$R_{y'} = r_{format}(y') + r_{acc}(y') + r_{refine}(y')$$

PASR

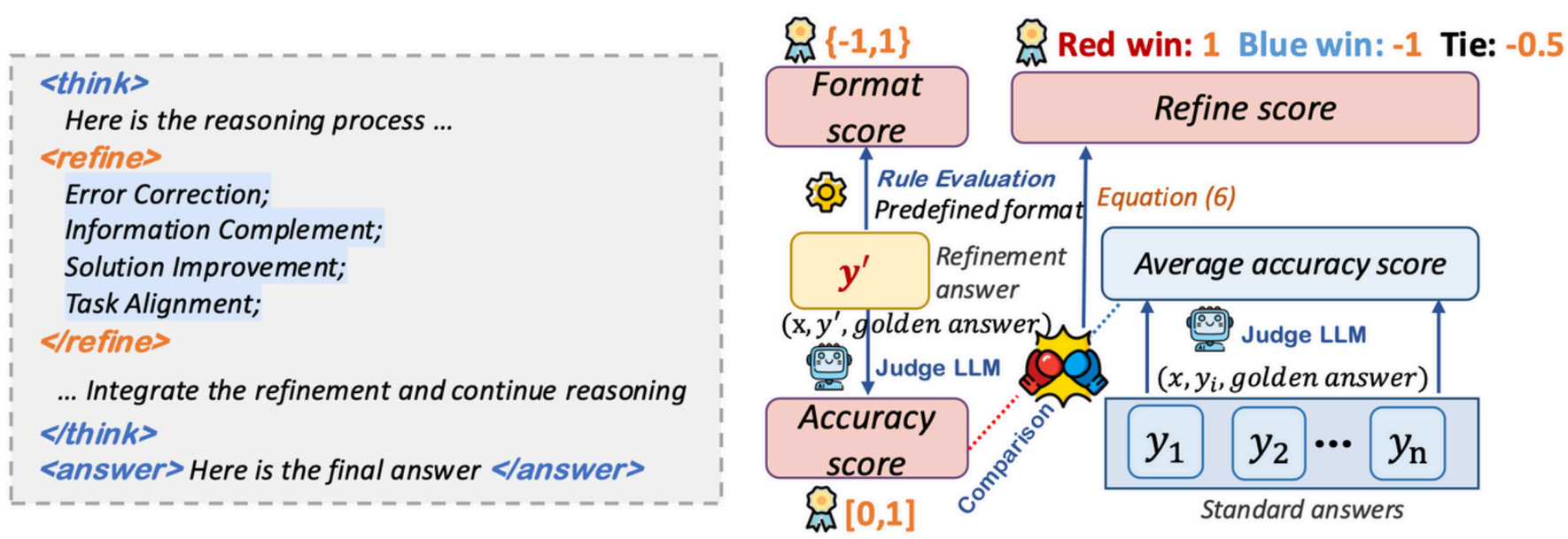

Figure 2: Answer format used in PASR (Left). Reward design for a generated answer y' during training (Right). The total reward is computed as the sum of the format score, accuracy score, and refinement score, as defined in Equation 7.

思考

在<think>...</think>内部增加<refine>...</refine>

1.昨天在写AWorld思考部分的内容时，还提到execution-guard和更常见的critic-refine的区别，前者是让guard在execution推理过程中就参与进来修正，后者是得到完整的response(reasoning trajectory)之后，critic再参与批评
2.本文的主动refine和著名的aha moment有区别吗？我觉得是有本质区别的，aha moment属于是llm涌现出来的，而本文是针对性的设计了prompt，让llm显式的去refine