arxiv.org/abs/2505.13379

Thinkless: LLM Learns When to Think

SFT+DeGRPO两阶段训练，让推理模型具备混合推理能力

开源代码：https://github.com/VainF/Thinkless

Gongfan Fang **Xinyin Ma** **Xinchao Wang***
National University of Singapore
gongfan@u.nus.edu, maxinyin@u.nus.edu, xinchao@nus.edu.sg

简介

本文提出了Thinkless框架，通过两阶段训练让推理(reasoning)模型能够针对数学任务，自动切换是用短回答还是长回答来求解，即具备混合推理能力。简单来说，Thinkless用两个特殊token <short>和<think>来作为response的第一个token，以此区分是短回答还是长回答，首先第一个阶段是sft训练，训练集包含了等量的(prompt, <short>短回答)和(promt, <think>长回答数据)，第二个阶段基于RLVR训练，为了更好的对两个特殊token进行优化，作者对GRPO做了两点改进：参考Dr. GRPO去掉reward标准差正则化以及分离控制/回答长度归一化。

背景

推理模型在复杂任务上有更好的表现，但是现实场景中，很多问题还是比较简单的，完全可以通过短回答直接解决，不需要长链推理。那么如何让推理模型能够根据任务(query)自动选择时进行短回答还是深思熟虑的长回答呢？本文提出了Thinkless方案。

实验设置

任务：MATH推理

- 框架：verl，实验对象：DeepSeek-R1-Distill-Qwen-1.5B，强化学习算法：DeGRPO
- RLVR reward function:

$$r(a, y^*, c) = \begin{cases} 1.0, & \text{if } c = \textit{<short>} \text{ and Extract-Answer}(a) = y^*, \\ 1.0 - \gamma, & \text{if } c = \textit{<think>} \text{ and Extract-Answer}(a) = y^*, \\ -1.0, & \text{if Extract-Answer}(a) \neq y^*, \end{cases}$$

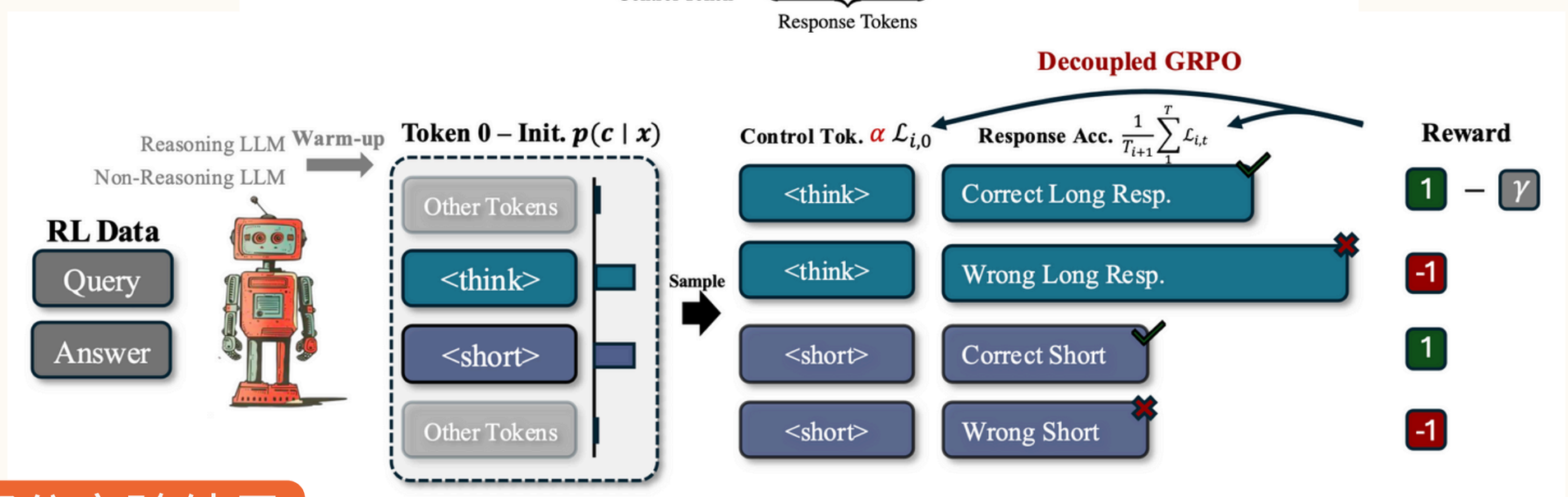
\gamma的作用是在保证正确率的前提下，引导模型尽量使用<short>模式回答，以降低推理成本。

训练流程

sft训练：这个阶段是为了让推理模型具备生成长回答和短回答的能力，为后续RL训练打下基础。为了创建sft训练集，找一批任务query，然后让一个更强的推理模型(DeepSeek-R1-671B)生成长回答a_{think}，让一个非推理模型(Qwen2.5-Math-1.5B-Instruct)生成短回答a_{short}，这样保证了长回答数据和短回答数据的数据量相同，然后构造长回答训练数据格式:(query, <think>, a_{think})，短回答训练数据格式: (query, <short>, a_{short})，然后做sft。

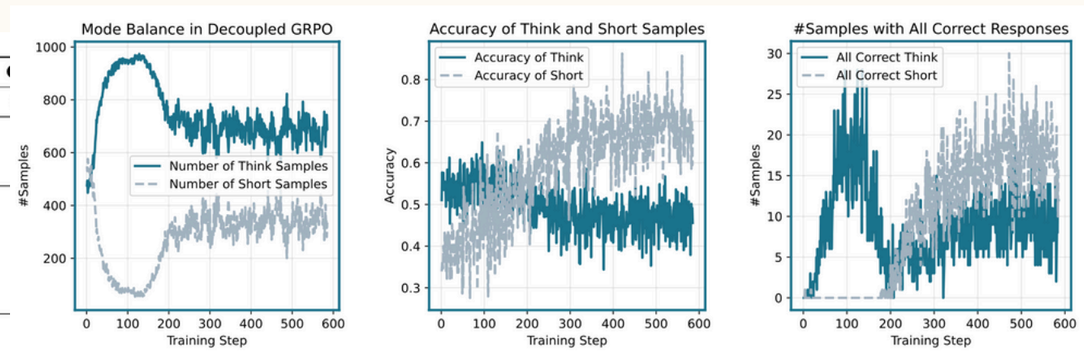
RL训练：经过sft之后，模型具备生成长回答和短回答能力了，但是还不会根据query难度自己选择用长回答还是短回答，为此用RL训练，作者对GRPO做了两点改进: 1）参考Dr. GRPO去掉reward标准差正则化; 2)为了加强对两个特殊token的优化，将特殊token和正式回答分离，分别作长度归一化。

$$\mathcal{J}_{\text{DeGRPO}}(\theta) = \mathbb{E}_{x,a_i}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\underbrace{\alpha\,\mathcal{L}_{i,0}(\theta)}_{\text{Control Token}} + \underbrace{\frac{1}{T_i}\sum_{t=1}^{T_i}\mathcal{L}_{i,t}(\theta)}_{\text{Response Tokens}} - \beta\,\mathbb{D}_{\text{KL}}\left[\pi_\theta(\cdot \mid x)\,\|\,\pi_{\text{ref}}(\cdot \mid x)\right]\Big)\right],$$



部分实验结果

Models	Type	AIME 2024		Minerva Algebra		Math-500		
		Pass@1	#Tokens (Think%)	Pass@1	#Tokens (Think%)	Pass@1	#Tokens (Think%)	Pass@1
DeepSeek-R1-1.5B	Base LLM	0.2800	18063	0.9577	3029	0.8608	5675	0.8347
Q-1.5B		0.0200	1300	0.7771	933	0.5168	855	0.7022
QMath-1.5B		0.1133	1128	0.9184	586	0.7604	721	0.8572
Merging-0.5 [34]	Short CoT	0.1333	8636	0.9292	834	0.7740	1524	0.8332
Merging-0.6 [34]		0.1733	10615	0.9321	1091	0.7900	3000	0.8381
Merging-0.7 [34]		0.1667	15854	0.9398	1834	0.8108	4347	0.8458
CoT-Valve $\alpha = 8$ [26]		0.2000	10692	0.8079	1903	0.7060	3723	0.7726
CoT-Valve $\alpha = 6$ [26]		0.1933	17245	0.9468	2656	0.8024	5167	0.7970
CoT-Valve $\alpha = 4$ [26]		0.2267	17722	0.9439	2965	0.8036	5820	0.8108
Router Random	Hybrid	0.1467	8093 (56.00%)	0.9211	1736 (49.28%)	0.7608	3096 (47.92%)	0.8205
Router Q-7B		0.1667	9296 (46.67%)	0.9250	795 (5.64%)	0.7948	2748 (25.00%)	**0.8587**
Thinkless		**0.2733**	7099 (100.00%)	**0.9459**	1144 (25.88%)	**0.8184**	2555 (51.56%)	0.8418



(b) The proposed Decoupled GRPO, with a U-shape learning curve.