arxiv.org/abs/2507.12759

Logit Arithmetic Elicits Long Reasoning Capabilities Without Training

开源代码：github.com/yunx-z/ThinkLogit

ThinkLogit: 将Proxy-tuning方法扩展到reasoning llm

Yunxiang Zhang* **Muhammad Khalifa** **Lechen Zhang** **Xin Liu**
Ayoung Lee **Xinliang Frederick Zhang** **Farima Fatahi Bayat** **Lu Wang**
Computer Science and Engineering
University of Michigan, Ann Arbor

简介

基于Proxy-tuning思想，本文提出ThinkLogit和ThinkLogit-DPO两种推理(Inference)阶段解码(decoding)方法，用于在不tuning base llm的前提下，激发其长链式推理(Long Chain-of-Thought)推理能力。核心做法是引入一个已经训练好的小reasoning模型，称为guider，通过logits arithmetic在推理过程中引导目标llm的输出。然后更进一步，使用Direct Preference Optimization(DPO)训练 guider，使其logit分布与目标llm更对齐，再应用proxy-tuning效果更好，这就是ThinkLogit-DPO。

背景

目前基于pre-training model构建reasoning llm的主流做法是RLVR， 训练成本还是有的，之前我们读过proxy-tuning论文，它不需要tuning 目标模型，而是只tuning一个小版本的llm，然后在inference阶段调整目标llm的logits就能逼近sft后的效果。本文将这一思想推广到reasoning llm场景，作者提出ThinkLogit方法，利用已经训练好的小reasoning llm来引导大规模的base model生成long CoT，从而无需tuning目标llm参数也能激活其潜在的reasoning 能力。

实验设置

- 实验对象：目标llm是Qwen2.5-32B，两个guider是R1-Distill-Qwen-1.5B和One-Shot-RLVR-1.5B
- 4个数学推理数据集：AIME2024, AIME2025, AMC23和MATH-hard

Model	# T.E.	# T.P.	AIME 2024	AIME 2025	AMC 23	MATH hard	Average
Large model baselines							
Qwen2.5-32B (***Target***)	-	-	14.6 / 40.0	8.3 / 26.7	57.2 / 90.0	50.8 / -	32.7 / 52.2
s1.1-32B	1K	32B	32.9 / 60.0	25.4 / 50.0	70.0 / 92.5	79.0 / -	51.8 / 67.5
R1-Distill-Qwen-32B	800K	32B	45.8 / 76.7	35.0 / 60.0	76.9 / 92.5	76.5 / -	58.6 / 76.4
Supervised Fine-tuned (SFT) LM as Guider							
R1-Distill-Qwen-1.5B (***Guider***)	-	-	16.2 / 33.3	18.8 / 33.3	51.2 / 80.0	47.7 / -	33.5 / 48.9
Target + THINKLOGIT	0	0	**22.5** / 50.0	19.2 / 36.7	62.2 / **95.0**	60.7 / -	41.2 / 60.6
Target + THINKLOGIT-DPO	10K	78M	22.1 / **60.0**	**21.7** / **46.7**	**63.7** / 90.0	**61.1** / -	**42.2** / **65.6**
Reinforcement Fine-tuned (RFT) LM as Guider							
One-Shot-RLVR-1.5B (***Guider***)	-	-	13.3 / 30.0	7.1 / 26.7	46.9 / 77.5	51.1 / -	29.6 / 44.7
Target + THINKLOGIT	0	0	**17.5** / **43.3**	**11.2** / **36.7**	**57.2** / **85.0**	**61.1** / -	**36.8** / **55.0**

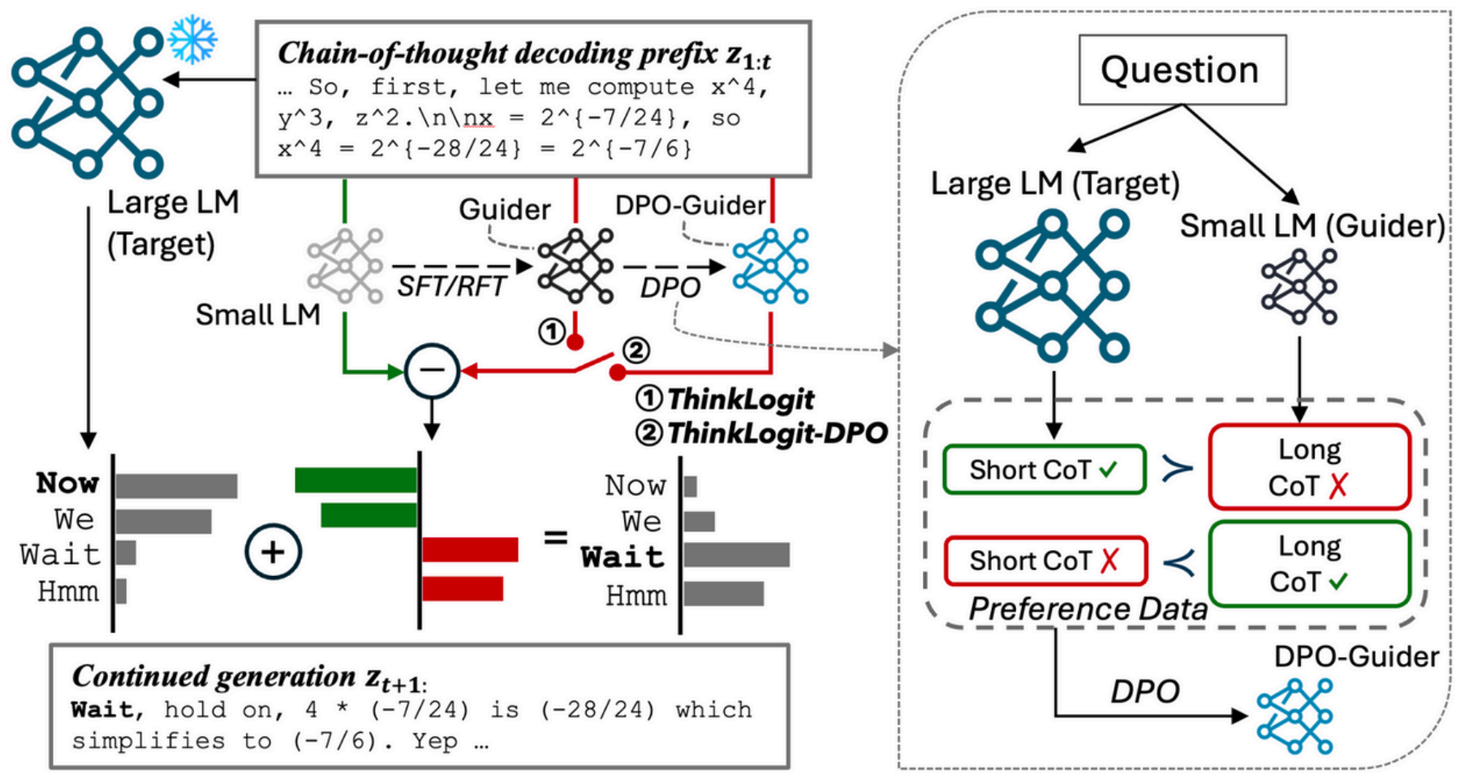
ThinkLogit和ThinkLogit-DPO

ThinkLogit就是proxy-tuning的应用。

ThinkLogit-DPO就是对已经tuning好的小reasoning llm继续DPO训练，重点是偏好数据集包含两类型数据：

- (x, 目标llm的short CoT正确，guider的long CoT错误)
- (x, 目标llm的short CoT错误，guider的long CoT正确)

构造这两类「大模型正确 vs 小模型错误」和「小模型正确 vs 大模型错误」数据，希望引导guider学习何时该跟随、何时该修正目标llm。



思考

本文延续了Proxy-tuning思想，同时作者假设pre-training 的base model已经具备推理能力，而sft/RLVR tuning后的reasoning llm只是激活了推理能力。目前很多工作都follow这种假设，见仁见智吧。

考虑到RL tuning相比于sft的难度和成本，或许近期会有更多的parameter-efficient RLVR tuning工作出现呢。

@机器爱学习