**Reshaping Large Language Models’ Meta-cognition via Reinforcement Learning**

**Motivation**

* 长CoT-RL训练带来了性能提升同时引入了**严重的幻觉问题和冗余tokens的问题**，严重损害了模型可用性
* 部分置信度矫正的方法引入了额外的分析和tokens，并且存在拒答率过高的问题，不能保证**线下场景的可用性**
* 有没有一种方法可以让模型：1. 推理性能获得提升 2. 不引入额外的tokens和分析，内化矫正模型的置信度和准确率的一致性，从而达到**幻觉控制**和**冗余tokens减少**的效果
* 受课程学习和openai *why LLM hallucinate* 启发，尝试通过**构造元认知任务来降低孤立知识点**的比例，提升模型**内部知识的一致性**，例如，在数学中：*判断哪种推理路径是好的比做出一道数学题往往更简单*
* 受tongyi-deepresearch, beyond tenturns and rstar2等任务的启发，抛弃复杂的算法设计，直接进行数据侧和参数侧的**scaling**

|  |  |  |
| --- | --- | --- |
|  |  |  |

**1017**

* ~~entropy设置成0~~
* ~~sample数量更多（8，16）~~
* 统计数学任务训练集235b得输出长度
* ~~训练前后的不同路径的比例~~
* 如何解耦推理路径和答案
* ~~训练一个评价器~~

**1022**

* end-to-end的 process\_reward 换成 子任务分解
* 评价器训练更细化
* 收集数据集
* 元认知任务的定义

**具体方法：**

* 直接scaling，构造大批量的数据
* 目标:1.数学推理更优的路径2.实现在冲突判别，错误信息识别任务中更好的性能提升
* 训练阶段：1. sft少量数据冷启动 2. rl做认知训练：推理路径判别任务，错误文档识别任务，多立场文档分类任务 3.下游任务rl训练
* 数学任务方法设计：判别更优路径 1. 设计四种路径，明显错误，隐蔽错误，正确但有修正，正确且简洁，选择不同路径，让模型判优 2. 让模型将两种答案改写的更接近。数据集内皆遵循题目从易到难
* rag任务设计：事实检测类任务，每个文档改写成一个错误文档，四个子任务：1. 判别改写后的文档和原文档哪个是错误的 2. 一个文档和另一个文档的改写文档哪个是错误的（太难） 3. 一个文档是事实性文档还是非事实性文档（太难） 4. 三个文档，分别判断是事实性还是非事实性，做对得全分，不做对不得分（太难）
* rag任务设计2：1. 判断某个query和它相关的几个文档，哪个对它的帮助更大（使用qwen235b构造golden\_answer然后使用4.1判断它之前说的对不对，一致率高达98%）2. fnc-1进行立场分类 3. 判别改写后的文档和原文档哪个是错误的
* 推理和认知的解耦：置信度和准确率的相关性

**预实验**

**实验设置**

* Grpo, clip\_ratio=0.2, qwen-3-4b base
* 训练集math, simple rl, 11499条

**实验结果：**

* 0926：第一版，使用token-level loss，生成长度过长，在快1 epoch时90%左右的case被截断，准确率在5steps时，即0.25 epoch时比较高,step5:80.2,step30:78.2
* 0928 8192：训了58步，clip太多了，我觉得也没意义了，最好的点是20:83 40:81
* 0928 16384: 长度太长了，改mini\_batch=16
* 0928\_4096\_sequence\_mean：还没把第一批数据训完就崩了，第二版加上kl loss
* 0928\_ours\_8192：加上kl\_loss
* 0928\_ours\_4096\_sequence\_mean：
* 0929\_ours\_4096\_sequence\_mean：把kl loss加上了，系数0.01
* 0930\_ours\_4096\_sequence\_mean:kl loss\*0.01，0.5format
* 0930\_4096\_sequence\_mean:把kl loss加上了，系数0.01，加上0.5格式损失,step5:77.8
* 0930\_4096:kl loss\*0.01，0.5format
* 0930\_ours\_4096:把kl loss加上了，系数0.01，加上0.5格式损失
* 0930\_8192:把kl loss加上了，系数0.01，加上0.5格式损失
* 1001\_ours\_4096\_sequence\_mean:kl loss\*0.01，0.5format,修复数据bug,step85,120:78.4，二阶段的认知训练太过简单了
* 1001\_ours\_4096:kl loss\*0.01，0.5format,修复数据bug，二阶段的认知训练太过简单了
* 1003\_4096:取消kl\_loss,0.5format
* 1003\_ours\_4096:取消kl\_loss，0.5format
* 1004\_4096:0.5format，课程学习4000条，1000条easy，medium,hard,very\_hard依次排列
* 1004\_4096\_ours: 0.5format，首先重做了math\_judge的数据，之前的二阶段，在这变成一阶段，并且把二阶段的数据重新洗了一遍，然后也是内部，从简单到容易排列
* 1009\_8192:发现entropy\_loss过高，越训练越大
* 1009\_8192\_ours:发现entropy\_loss过高，越训练越大
* 1013\_8192\_14b: batch\_size 128 tbs 16 entropy\_coef 1e-2
* 1013\_8192\_14b\_ours: batch\_size 128 tbs 16 entropy\_coef 1e-2
* 1014\_8192\_14b\_ours: entropy\_coef 1e-3
* 1015\_8192\_4b\_ours:entropy\_loss降到1e-3
* 1014\_8192\_4b:entropy\_loss降到1e-3
* 1016\_2048\_4b\_rag: entropy\_loss 1e-3,f1 reward
* 1017\_2048\_4b\_rag：entropy\_loss 1e-2
* 1017\_2048\_4b\_rag\_ours:构造了一个四阶段任务，一起训练, entropy\_loss 1e-2
* 1017\_4096\_4b\_ours: entropy\_loss置0
* 1017\_8192\_14b\_ours：entropy\_loss置3e-3
* 1018\_4096\_4b: entropy\_loss置0
* 1018\_2048\_4b\_rag：entropy\_loss 0
* 1018\_2048\_4b\_rag\_ours: entropy\_loss 0，数据有点小问题，把下游任务训练放前头了，重新训一般
* 1020\_8192\_14b\_ours：entropy\_loss置为0
* 1022\_4096\_14b\_rag: entropy\_loss 0,mini\_batch\_size 32
* 1022\_4096\_14b\_rag\_ours: entropy\_loss 0,mini\_batch\_size 32
* 1022\_4096\_4b\_rag\_ours: entropy\_loss 0,mini\_batch\_size 64
* 1022\_8192\_14b：entropy\_loss置为0，莫名其妙训崩了
* 1023\_8192\_14b\_1: entropy\_loss置为0
* 1023\_8192\_14b\_2:entropy\_loss置为1e-5
* 1024\_4096\_14b\_mixed\_rag\_math:entropy\_loss 0,mini\_batch\_size 32
* 1024\_4096\_4b\_mixed\_rag\_math: entropy\_loss 0,mini\_batch\_size 32
* 1027\_4096\_14b\_rag\_ours: 之前的实验有问题，数据的位置出错了，重跑！先过两遍认知再过两篇下游任务，不好!
* 1027\_4096\_4b\_rag\_ours: 之前的实验有问题，数据的位置出错了，重跑！先过两遍认知再过两篇下游任务，不好!
* 1029\_8192\_4b\_sftmath: dapo数据：共5000条sft100步，两个epoch,训练了30步,reward直接崩了，感觉是4b吞不下这个sft，或者sft不充分
* 1030: 数学任务统一采用llm\_as\_a\_judge做为reward保证reward信号的准确



* 1029\_8192\_14b\_sftmath\_rl: 从sft295开始
* 1027\_4096\_4b\_rag\_ours:先认知再下游，一共过两遍，交替进行
* 1030\_2048\_4b\_rag\_ours: 用了新的认知数据
* 1030\_4096\_14b\_rag\_ours: 用了新的认知数据
* 1030\_8192\_14b\_sftmath\_ours: 从sft295开始，做认知训练rl
* 1031\_2048\_4b\_rag\_ours: 使用mixed\_rag\_3，前面三个认知训练任务分开，不放在一起训练
* 1031:认知任务立场检测有bug:disscuss太多了！四个类别严重不平衡！disagree全部用上，剩下三类各取2000条实验
* 1031\_1\_2048\_4b\_rag\_ours: 修复上述bug
* 1031\_2\_2048\_4b\_rag\_ours: 修复上述bug，并使用stance在前的数据
* 1031\_4096\_14b\_rag\_ours: 修复上述bug

**key\_insights:**

* 加上kl之后貌似更稳定
* 4096看下来是最好的
* 16384感觉也没啥用
* 4.1mini llm\_as\_a\_judge的信号还没有路径评价模型准
* 本地vllm测还不如直接val结果高
* 断点继续训练结果往往都有问题
* 16384prompt+生成长度最好用16，否则可以用32
* 1万条数据4bmodel 3epochs训50h,0.5epoch效果最好
* kl\_loss和format不能同时加，可能效果会变差，考虑把kl\_loss去掉
* 经过认知训练后模型的输出更加稳定，长度也更加统一

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| version | 参数 | step | kl | format | token | mean\_mode | max\_tokens | entropy\_loss | math500 | olymid | minervamath | amc22 | amc23 | avg | AIME2024 | AIME2025 | avg | FELM  ba/a0/a1 |
| base | 4b | - | - | - | - | - | 8192 | 0.01 | 75.1 | 44.1 | 49.4 | 39.5 | 56.3 | 52.3 |  |  |  | 55.2/27.9/82.6 |
| 0928-token8192-steps20 | 4b | 20 | 0 | 0 | 8192 | tokenmean | 8192 | 0.01 | 81.7 | 53.7 | 60.8 | 48.8 | 62.5 | 61.5 |  |  |  |  |
| 0930-token4096-sequencemean-steps10 | 4b | 10 | 0.01 | 0.5 | 4096 | sequencemean | 8192 | 0.01 | 77.8 | 49.0 | 60.5 | 41.9 | 58.8 | 57.6 |  |  |  |  |
| 0930-token4096-sequencemean-steps40 | 4b | 40 | 0.01 | 0.5 | 4096 | sequencemean | 8192 | 0.01 | 77.5 | 44.7 | 63.1 | 39.5 | 51.3 | 55.2 |  |  |  |  |
| 1001-token4096-ours-sequencemean-steps115 | 4b | 120 | 0.01 | 0.5 | 4096 | sequencemean | 8192 | 0.01 | 78.4 | 46.0 | 63.2 | 41.9 | 62.5 | 58.4 |  |  |  |  |
| 0930-token4096-steps5 | 4b | 120 | 0.01 | 0.5 | 4096 | tokenmean | 8192 | 0.01 | 75.8 | 47.7 | 56.6 | 41.9 | 63.8 | 57.1 |  |  |  |  |
| 1001-token4096-ours-steps90 | 4b | 90 | 0.01 | 0.5 | 4096 | tokenmean | 8192 | 0.01 | 73.6 | 41.2 | 58.1 | 41.9 | 45 | 52.0 |  |  |  |  |
| 1004-token4096-steps35 | 4b | 35 | 0 | 0.5 | 4096 | tokenmean | 8192 | 0.01 | 81.3 | 53.3 | 60.3 | 37.2 | 62.5 | 58.9 | 10 | 23.3 | 16.7 | 57.4/33.3/81.2 |
| 1004-token4096-ours-steps75 | 4b | 75 | 0 | 0.5 | 4096 | tokenmean | 8192 | 0.01 | 82.1 | 52.0 | 68.0 | 53.5 | 57.5 | 62.6 | 20 | 15.0 | 17.5 | 58.1/38.1/78.2 |
| 1003-token8192-ours-steps40 |  | 40 |  | 4096 | 4096 | tokenmean | 8192 | 0.01 | 72.4 | 42.1 | 60.1 | 27.9 | 56.3 |  | 20 | 3.3 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1003-token4096-steps90 |  | 90 |  | 4096 | 4096 | tokenmean | 8192 | 0.01 | 77.5 | 43.4 | 66.0 | 34.9 | 57.5 |  | 10 | 11.7 |  |  |
| 1009-token8192-ours-steps35 |  | 35 | 0 | 8192 | 8192 | tokenmean | 8192 | 0.01 | 82 | 55.4 | 68.2 | 46.5 | 52.5 | 60.9 | 23.3 | 13.3 | 18.3 | 60.4/40.8/80.0 |
| 1009-token8192-steps40 |  | 40 | 0 | 8192 | 8192 | tokenmean | 8192 | 0.01 | 75.2 | 42.1 | 67.3 | 37.2 | 47.5 | 53.9 | 13.3 | 10 | 11.6 | 59.2/34.7/83.6 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1015-token8192-steps30-ours | 4b | 30 | 0 | 8192 | 8192 | tokenmean | 8192 | 1e-3 | 79 | 47.6 | 62.9 | 44.2 | 47.5 | 56.2 | 13.3 | 20 | 16.7 | 59.4/42.2/76.6 |
| 1014-token8192-steps35 | 4b | 35 | 0 | 8192 | 8192 | tokenmean | 8192 | 1e-3 | 79.8 | 45.3 | 63.6 | 41.9 | 48.9 | 55.9 | 10 | 16.6 | 13.3 | 53.61/21.8/85.5 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1014-token8192-ours-steps40-14b | 14b | 40 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-3 | 83.3 | 51.3 | 66.5 | 54.7 | 67.5 | 64.7 | 10 | 13.3 | 11.7 |  |
| 1013-token8192-ours-steps65-14b | 14b | 75 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-2 | 86.2 | 57.7 | 70.0 | 53.5 | 67.5 | 67.0 | 23.3 | 15 | 19.2 | 63.9/51.7/76.1 |
| 1013-token8192-steps65-14b | 14b | 65 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-2 | 84 | 52.1 | 75.9 | 53.5 | 72.5 | 67.6 | 13.3 | 13.3 | 13.3 | 60.0/36.1/83.9 |
|  |  |  |  |  |  |  |  |  | math500  4r major | olymid  4r major | minervamath  4r major | amc22  16r major | amc23  16r major | avg  major | AIME2024  16r major | AIME2025  16r major | avg  major |  |
| qwen-14b-base | 14b |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1020-token8192-ours-step170-14b | 14b | 170 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-5 | 84.4 | 53.2 | 70.2 | 46.5 | 67.5 | 64.4 | 16.7 | 20 | 18.3 | 62.6/42.9/82.3 |
| 1023\_1-token8192-step85-14b | 14b | 85 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-5 | 83.8 | 48.4 | 71.3 | 51.2 | 57.5 | 62.4 | 16.7 | 20.0 | 18.3 | 61.9/42.2/81.6 |
| 1023\_2-token8192-step20-14b | 14b | 20 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-5 | 85.2 | 56.8 | 52.9 | 53.5 | 67.5 | 63.0 | 20.0 | 20.0 | 20.0 | 57.5/20.4/94.5 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1029\_8192\_14b\_sftmath | 14b |  |  |  |  |  | 8192 | 1e-5 | 89.2 | 59.2 | 73.2 | 58.1 | 80.0 | 72.0 | 40.0 | 40.0 | 40.0 |  |
| 1029\_8192\_14b\_sftmath\_rl\_step | 14b | 20 | 0 | 0.5 | 8192 | tokenmean | 8192 | 1e-5 |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1017-4b-token4096-ours-step75 | 4b | 75 | 0 | 0.5 | 4096 | tokenmean | 8192 | 1e-5 | 78.4 | 46.4 | 60.6 | 41.8 | 65.0 | 58.4 | 10.0 | 10.0 | 10.0 | 56.5/37.4/75.6 |
| 1018-4b-token4096-step45 | 4b | 45 | 0 | 0.5 | 4096 | tokenmean | 8192 | 1e-5 | 78.6 | 44.8 | 62.1 | 39.5 | 52.5 | 55.1 | 13.3 | 6.67 | 10.0 | 53.9/27.2/80.5 |

* 4b下4096>8192，清洗过的数据>未清洗数据，entropy\_loss0>0.01>0.001
* 14b下，entropy\_loss0.01>0.001>1e-5>0，越低训练更稳定，
* 14b下，entropy置零训练明显更稳定也更好
* 在4b模型下，不同tokens和方法下我们的方法与直接训练的方法在域内任务上显著高于直接训练的方法
* 在14b模型size下，我们的方法与直接训练的方法性能相当，后续通过精细化的数据收集和数据构造试图解决这个问题
* 在更难的数学问题和事实检测任务上，我们的方法表现出了强大的优势，表明我们的方法在**幻觉控制上**和**更难问题下推理路径**的选择上更具有优势。
* 14b上未见我们claim的**高效**
* rag任务很失败啊，感觉这任务不行啊，认知任务设计的不好！

|  |  |
| --- | --- |
| qwen3-14b-ours-steps65  aime2024    math500 | qwen3-14b-steps65  aime2024    math500 |

|  |  |
| --- | --- |
| qwen3-4b-ours-steps75  aime2024    math500 | qwen3-4b-steps35  aime2024    math500 |

|  |  |
| --- | --- |
| Math500 ours  qwen-3-4b-base-math-grpo1004-token4096-ours-steps75.cot | Math500 baseline  qwen-3-4b-base-math-grpo1004-token4096-steps35.cot |

* 14b的amc23也差不多，以及你分类器太不鲁棒了，要重新训练啊！

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| version | 参数 | step | kl | format | token\_train | mean\_mode | entropy\_loss | max\_tokens\_gen | overall\_acc  r8 major | no\_conf  r8 major | outdated\_info  r8 major |
| base-4b | 4b | - | - | - | - | - | - | 2048 | 74.0 | 82.6 | 51.6 |
| 1018-token2048-steps60 | 4b | 60 | 0 | 0.5 | 2048 | tokenmean | 0 | 2048 | 74.6 | 82.5 | 54.2 |
| 1024-mathrag-token4096-steps45 | 4b | 45 | 0 | 0.5 | 4096 | tokenmean | 0 | 2048 | 74.9 | 83.2 | 53.2 |
| 1022-token2048-steps175  -ours | 4b | 175 | 0 | 0.5 | 2048 | tokenmean | 0 | 2048 | 75.8 | 83.6 | 55.6 |
| 1028-token2048-steps180  -ours | 4b | 180 | 0 | 0.5 | 2048 | tokenmean | 0 | 2048 | 75.8 | 83.9 | 54.8 |
| 1022-token2048-steps175  -ours-raw\_prompt | 4b | 175 | 0 | 0.5 | 2048 | tokenmean | 0 | 2048 | 74.4 | 83.9 | 50.0 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| base-14b | 14b | - | - | - | - | - | - | 2048 | 78.0 | 85.1 | 59.5 |
| 1018-token4096-steps80 | 14b | 80 | 0 | 0.5 | 4096 | tokenmean | 0 | 2048 | 79.8 | 87.0 | 61.3 |
| 1024-mathrag-token4096-steps125 | 14b | 125 | 0 | 0.5 | 4096 | tokenmean | 0 | 2048 | 81.2 | 88.2 | 62.9 |
| 1022-token4096-steps70  -ours | 14b | 70 | 0 | 0.5 | 4096 | tokenmean | 0 | 2048 | 80.3 | 86.9 | 62.9 |

**Truthfulqa**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| size | 任务 | Model | r8\_avg\_em | +/- |
| 4b | - | 4b-base | 52.0 | 0 |
| 4b | rag | qwen3-4b-base-rag-grpo20251018-token2048-steps60 | 51.0 | -1.0 |
| 4b | math-rag | qwen3-4b-base-math-rag-grpo20251024-token4096-steps45 | 50.2 | -1.8 |
| 4b | rag-ours | qwen3-4b-base-rag-ours-grpo20251022-token2048-steps175 | 56.7 | +4.7 |
| 14b | - | 14b-base | 64.6 | 0 |
| 14b | math | qwen3-14b-base-math-grpo14b-20251023\_1-token8192-steps85 | 67.7 | +3.1 |
| 14b | sftmath\_rl | qwen3-14b-base-sftmath-grpo14b-20251030-token8192-steps30 | 65.4 | +0.8 |
| 14b | math-ours | qwen3-14b-base-math-grpo14b-20251020-token8192-ours-steps170 | 68.9 | +4.3 |

**数学路径更优判别：**

* 判别模型：8b训不起来，换14b吧，数据也有问题
* 任务构造：4b和14b分别在math500，minervamath, AIME25里面的rollout，四个模型：  
  qwen3-14b-base-math-grpo14b-20251020-token8192-ours-steps170.cot，qwen3-14b-base-math-grpo14b-20251023\_1-token8192-steps85.cot，qwen3-4b-base-math-grpo4b-20251018-token4096-steps45.cot,qwen3-4b-base-math-grpo4b-20251017-token4096-ours-steps75.cot，一共12个文件，定义如下四个任务：1. 每个题目每个模型自己的rollout里面选择两个相同答案和两个不同答案的路径 2. 每个模型模型和自己同参数的模型，每个题目的rollout内部随机选择一个rollout，形成一个pair 3.每个模型和自己同类型的模型（ours和ours）每个题目的rollout内部随机选择一个rollout，形成一个pair
* 使用qwen235b判优，8个rollout，必须有5个以上达成共识，才算golden标签，**加了个confidence，结果全是high，一致率很低的问题也全是high**

**rag任务：**

**立场检测子任务：**

|  |  |
| --- | --- |
| 14b-base-acc | 4b-base-acc |