arxiv.org/abs/2504.02495

Inference-Time Scaling for Generalist Reward Modeling

DeepSeek-GRM: 考虑Inference-Time Scalability的支持多输入格式的Reward Model

Zijun Liu[1,2†*], **Peiyi Wang**[1*], **Runxin Xu**[1], **Shirong Ma**[1], **Chong Ruan**[1], **Peng Li**[3], **Yang Liu**[2,3], **Yu Wu**[1]

[1]DeepSeek-AI, [2]Dept. of Computer Sci. & Tech., Tsinghua University,
[3]Institute for AI Industry Research (AIR), Tsinghua University

zj-liu24@mails.tsinghua.edu.cn, wangpeiyi9979@gmail.com

简介

本文提出DeepSeek-GRM，用于探索Inference-time scalability的通用领域RM。1) 为适配不同输入格式，DeepSeek-GRM采用pointwise结构，即为每一个response都生成一个reward值，这样就不需要关系一个query带有几个response了；2) 针对RM训练，提出两阶段训练方法SPCT(Self-Principled Critique Tuning，包含Rejective Fine-Tuning(RFT)和RLVR，RFT阶段作为初始化，希望GRM能适应输入格式，并且知道要生成principles、critiques和reward值，RLVR阶段用GRPO算法训练；3) 得到DeepSeek-GRM后，本文探索了inference-time scalability，简单说就是多采样 + meta RM 过滤后的投票机制，进一步提升GRM效果。

背景

RL scaling在llm post-training中日益重要，因此Reward Model(RM)的地位也水涨船高。如果能实现真正的通用领域RM，无疑将极大推动llm的发展。问题是：1)格式适配: RM若要适应不同领域和任务，需要支持单response、成对response、多个response 等输入格式，如何设计RM? 2)如何训练RM? 3）如何考虑RM的inference-time scalability

实验设置

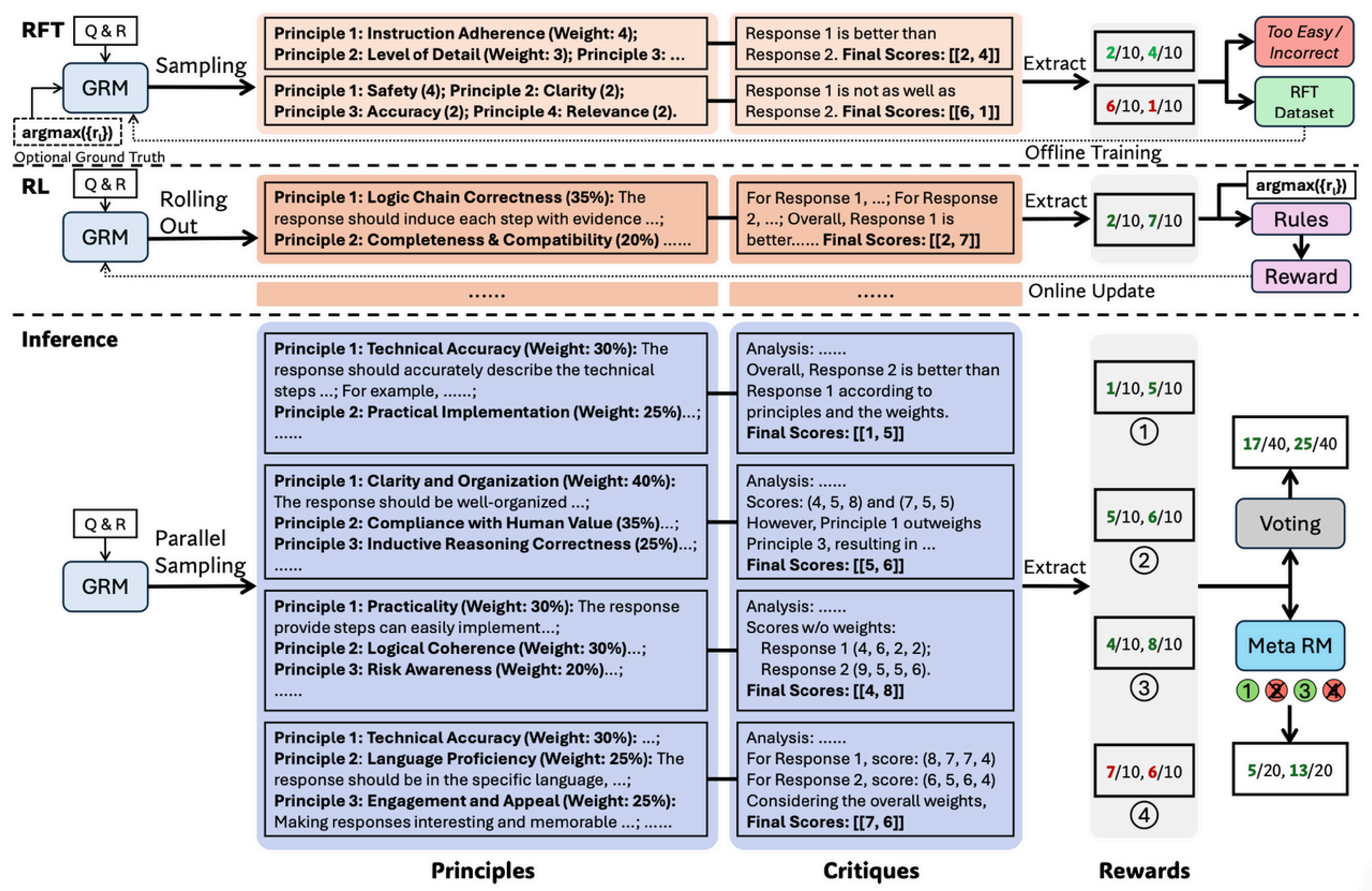
- GRM模型：Gemma-2-27B
- RM训练集：私有训练集，大约1250K
- GRPO训练时去掉了format reward
- 考虑principle的reward赋值：DeepSeek-GRM面对每组(query, response)先生成principles(来自Anthropic Constitutional AI的概念)，再生成critique和reward值

$$\{p_i\}_{i=1}^{m} \sim p_\theta\left(x, \{y_i\}_{i=1}^{n}\right), \quad \mathcal{R} = \mathbf{C} \sim r_\theta\left(x, \{y_i\}_{i=1}^{n}, \{p_i\}_{i=1}^{m}\right),$$

SPCT两阶段训练

训练用的RM数据格式是一个prompt对应N个response，N>=1，同时有一个index j指明哪条response最合理，index j作为ground truth。

1. RFT，由于训练集RM数据不包含principle和critique，因此先让llm生成，再根据生成结果过滤，因此用的是RFT而不是sft
2. 基于GRPO的RLVR



思考

由于对这篇论文期待比较高，读完之后其实是略微失望的（我希望是我没有理解透），简单说，这篇论文做了两个事情：训练一个适用多领域的GRM，有了GRM后，考虑在inference时如何inference-time scalability。为什么有些失望呢，第一，看到通用领域的reward model，我第一想法是估计要从人类为response打标签难度入手，提出了某种自动标注preference data的方案，甚至不标注做隐式训练，结果训练集是有ground truth的，或许本文的GRM的G更多的是指能接受多种输入格式的数据吧，第二是关于principle，读前面关于让GRM先生成principle再生成critique和reward值，绕这么一大圈，我以为是要做无ground truth reward训练呢，结果是监督数据集，再看消融实验，我并不觉得加了principle就有多么重要，当然关于2个点的提升到底大不大，因人而异吧；第三是inference-time scaling，也只是parallel sampling + voting，略微平平无奇。当然了，本文也没开源代码。

以上吐槽并不影响本文还是一篇质量上乘的reward model论文。