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 Xu¹, Shirong Ma¹, Chong Ruan¹, Peng Li³, Yang Liu^{2,3}, Yu Wu¹

¹DeepSeek-AI, ²Dept. of Computer Sci. & Tech., Tsinghua University,

³Institute for AI Industry Research (AIR), Tsinghua University

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

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 Al的概念),再生成critique和reward值

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

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IIm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

Principle 1: Instruction Adherence (Weight: 4); Extract 2/10, 4/10 Response 2. Final Scores: [[2, 4]] Principle 2: Level of Detail (Weight: 3); Principle 3: . Sampling rinciple 1: Safety (4); Principle 2: Clarity (2); KFI **6**/10, **1**/10 Response 2. Final Scores: [[6, 1]] Principle 3: Accuracy (2); Principle 4: Relevance (2). Dataset argmax({r_l}) Offline Training Optional Ground Truth RLQ&R argmax({r_i}) Rolling Principle 1: Logic Chain Correctness (35%): The For Response 1, ...; For Response 2, ...; Overall, Response 1 is Rules response should induce each step with evidence ...; 2/10, 7/10 **GRM** Principle 2: Completeness & Compatibility (20%) better Final Scores: [[2, 7]] Reward Online Update Inference Principle 1: Technical Accuracy (Weight: 30%): The Analysis: Overall, Response 2 is better than response should accurately describe the technical **1**/10, **5**/10 steps ...; For example,; Response 1 according to principles and the weights. Principle 2: Practical Implementation (Weight: 25%)...; 1 Final Scores: [[1, 5]] 17/40, 25/40 Principle 1: Clarity and Organization (Weight: 40%): Analysis: The response should be well-organized ...; Scores: (4, 5, 8) and (7, 5, 5) 5/10, 6/10 Q&R Principle 2: Compliance with Human Value (35%)...; However, Principle 1 outweighs Voting **Parallel** Principle 3: Inductive Reasoning Correctness (25%)...; Principle 3, resulting in ... 2 Sampling Final Scores: [[5, 6]] Extract **GRM** Principle 1: Practicality (Weight: 30%): The response Analysis: ... provide steps can easily implement ...; Scores w/o weights: 4/10, 8/10 Meta RM Principle 2: Logical Coherence (Weight: 30%)...; Response 1 (4, 6, 2, 2); Principle 3: Risk Awareness (Weight: 20%)...; Response 2 (9, 5, 5, 6). 3 1 2 3 % Final Scores: [[4, 8]] Principle 1: Technical Accuracy (Weight: 30%): ...; Analysis: ... Principle 2: Language Proficiency (Weight: 25%): The For Response 1, score: (8, 7, 7, 4) 5/20, 13/20 7/10, 6/10 response should be in the specific language, ...; For Response 2, score: (6, 5, 6, 4) Principle 3: Engagement and Appeal (Weight: 25%): Considering the overall weights, 4 Making responses interesting and memorable ...; .. Final Scores: [[7, 6]] **Principles** Critiques Rewards

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。