arxiv.org/abs/2507.18624

Checklists Are Better Than Reward Models For Aligning Language Models

RLCF：将checklist融入DPO

Vijay Viswanathan ♡ Yanchao Sun♣ Shuang Ma♣* Xiang Kong♣
Meng Cao♣ Graham Neubig♡ Tongshuang Wu♡
♡ Carnegie Mellon University ♣Apple

计划开源

简介

本文提出RLCF(Reinforcement Learning from Checklist Feedback)，简单来说，基于checklist对response打分构造preference data然后用作DPO训练。注意并不是根据checklist对response打分作为reward值，而是通过llm自动从指令中提取checklist，逐项评估response质量，并以此构建“chosen vs rejected”的preference data，然后DPO训练。

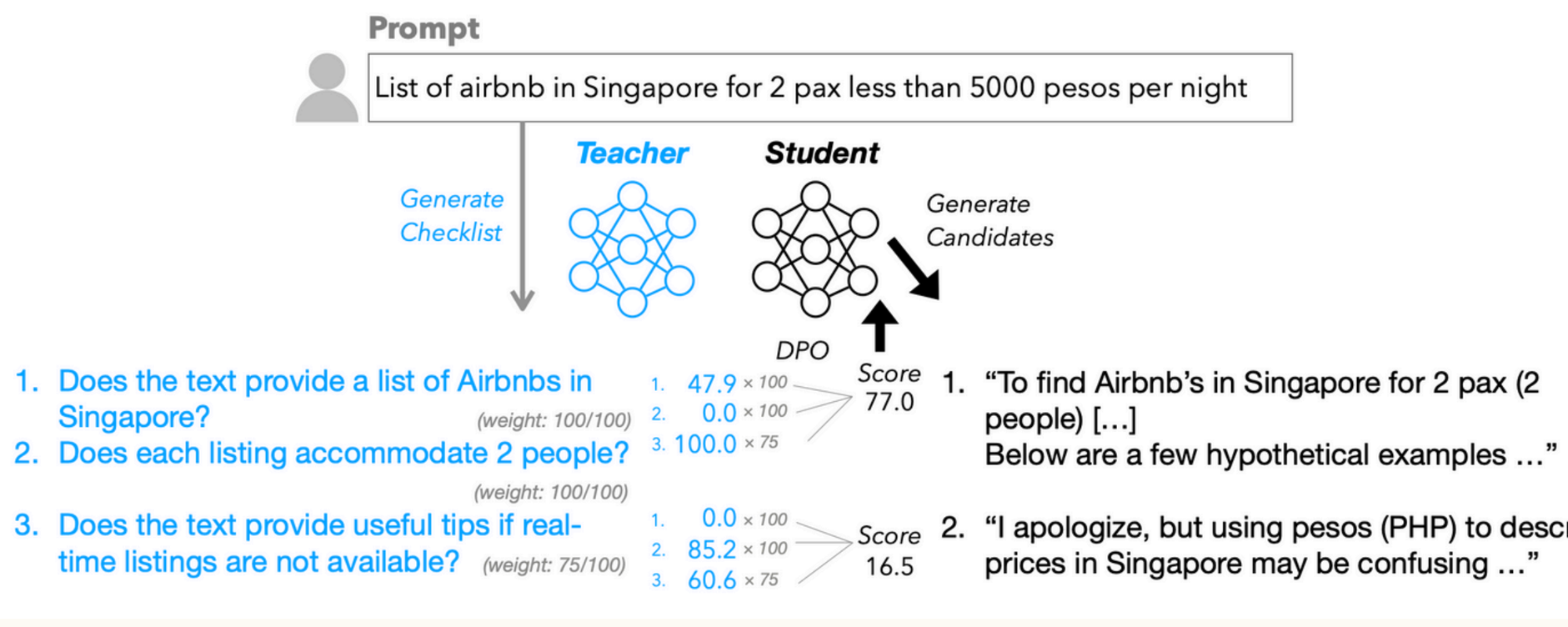
背景

前面我们读过一篇checklist论文，作用是评估llm的response，由于能给出细粒度的可解释的评估，当时作者进一步用checklist分数作为反馈对response做refinement。本文更进一步，将checklist用于RL训练阶段，但是要注意虽然checklist提供了具体的 reward分数，但本文并没有用GRPO、PPO等RL算法，而是选择无需reward值的DPO算法。

实验设置

- 实验对象：Qwen2.5-7B 和 Qwen2.5-7B-Instruct
- 生成checklist以及打分的模型：Qwen2.5-72B-Instruct
- rl框架：OpenRLHF
- 偏好数据集：WildChecklists, 130K条数据
- 为了防止reward hacking，RLCF在每条checklist中统一加入了一条通用约束项(universal requirement)，明确要求response必须直接回答instruction，避免冗余或跑题内容，并匹配指令语境下的语气与风格。此外，在构造偏好对时，仅保留打分差异最大的前40%样本

RLCF



思考

前面我们读过的TICK和STICK已经提出了用checklist分数作为反馈来refine llm的response，本文更进一步，将checklist分数融入到强化学习阶段，只不过想不明白为什么不直接作为reward值用GRPO等rl算法训练呢？