arxiv.org/abs/2502.00640

CollabLLM: 从单轮response优化到多轮协作优化

CollabLLM: From Passive Responders to Active Collaborators

Shirley Wu [1] Michel Galley [2] Baolin Peng [2] Hao Cheng [2] Gavin Li [1] Yao Dou [3] Weixin Cai [1]
James Zou [1] Jure Leskovec [1] Jianfeng Gao [2]

http://aka.ms/CollabLLM

简介

开源代码：https://github.com/Wuyxin/collabllm

本文提出CollabLLM，一个面向多轮对话(multiturn conversation)场景的llm tuning框架，目的是将llm从“被动响应者”转变为“主动协作者”。用rlvr的话来说，就是不再通过比较response与ground truth之间的差异来计算reward值，而是根据这条response对后续conversation的影响构造reward值，问题是如何计算这个理论上的后期影响reward值呢？作者提出了multiturn-aware(多轮感知) reward去近似，通过llm模拟未来几轮的对话(forward sampling)，然后计算多种客观的和主观的指标作为reward值，然后用强化学习训练。本文计算multiturn-aware reward的方法是通过prompt llm实现的，不需要tuning，因此Collabllm还可以用于创建包含reward值的训练数据，这无疑扩大了可以应用的范围。

背景

本文属于multiturn conversation方向的工作，我们在和llm交流的时候，很多情况下难以在第一个query中就将自己的需求或目的描述清楚，往往需要和llm沟通好几轮才可能解决问题。目前llm的训练范式(比如sft、rlhf、rlvr)基本上都是在优化单轮(single-turn)的response，这样得到的llm倾向于针对用户query直接生成答案，而不是引导式地与用户协同多沟通来解决问题，作者将单轮优化模式称为被动响应(passive responser)，虽然在静态的benchmark上表现越来越好，但实际多轮交互中一旦不能领会用户意图往往就会沟通很低效。

作者认为评价一个response是否好，不应该局限在对当前query的响应，而是应该从multiturn conversation全局看，是否能在后续的对话中带来高质量的互动从而解决用户问题提升用户体验，这种模式称为主动协作(active collaborator)，这就是本文CollabLLM的核心出发点：对llm response的reward不再只看当下是否有用，而是看这条response对后续整体对话的长期影响。

CollabLLM

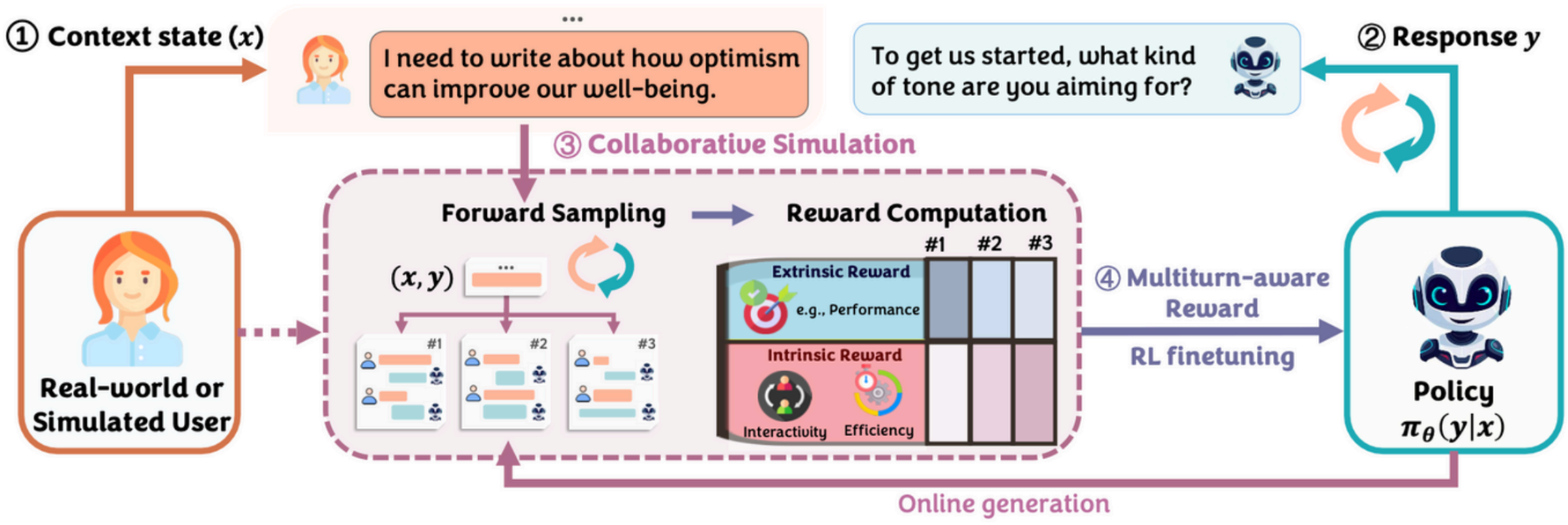

Figure 1: CollabLLM Framework: Given a context ①, the model generates a response ② to maximize long-term collaboration gains, termed *Multiturn-aware Rewards* (MR). During training, MRs are estimated via ③ collaborative simulation, which forward-samples conversations with simulated users. Finally, ④ reinforcement fine-tuning is applied using the MRs.

核心是如何计算reward值，评估一条response对后期对话的影响。作者用forward sampling + multiturn-awared reward去模拟。

思考

这篇论文是小红书上的朋友推荐的，我查了一下发现是ICML 2025的Outstanding Paper，因此在阅读过程中难免会受到一些先入为主的影响。论文关注的是多轮对话(multiturn conversation)层面的优化，相比于传统的单轮response优化，要更加high level一些，毕竟讨论的是一个更大的问题。作者提出的方法，我觉得也很有想法，至于它的实际效果以及对这个领域未来的影响，已经超出了我的评价能力。总的来说，是一篇非常不错的论文。