arxiv.org/abs/2507.01352

Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy

第二代 skywork open reward model

Chris Yuhao Liu Liang Zeng Yuzhen Xiao Jujie He Jiacai Liu Chaojie Wang
Rui Yan Wei Shen Fuxiang Zhang Jiacheng Xu Yang Liu Yahui Zhou

2050 Research, Skywork AI

https://huggingface.co/Skywork
https://modelscope.cn/organization/Skywork

简介

本文提出Skywork-Reward-V2，是Skywork开源的第二代reward model，作者从构建大规模高质量偏好(preference)数据集入手，首先构造了SynPref-40M(4000万对偏好数据)数据集，然后在此基础上训练了一系列从0.6B到8B的reward model。本文的重点是如何构造SynPref-40M数据集，为此作者设计了两阶段数据标注流程：第一阶段通过严格的人工验证结合llm judge辅助标注，以迭代方式不断获取高质量偏好数据；第二阶段利用训练好的reward model自动在海量偏好数据中筛选并用llm修正标注结果，实现自动规模化扩展。

注意：数据集并未开源

背景

虽然RLVR for llm reasoning已经成为研究主流，但是对于人类偏好(preference)，还是很难用规则的方式来设计有效的reward function，偏好建模仍离不reward model，而目前开源社区中的reward model效果还不能令人满意。

如何得到高质量的reward model？可以是选择更好的base model，也可以是设计更高超的训练技巧，而本文则更直接，从偏好训练集入手，创建了包含4千万偏好数据的SynPref-40M，然后训练reward model并开源。

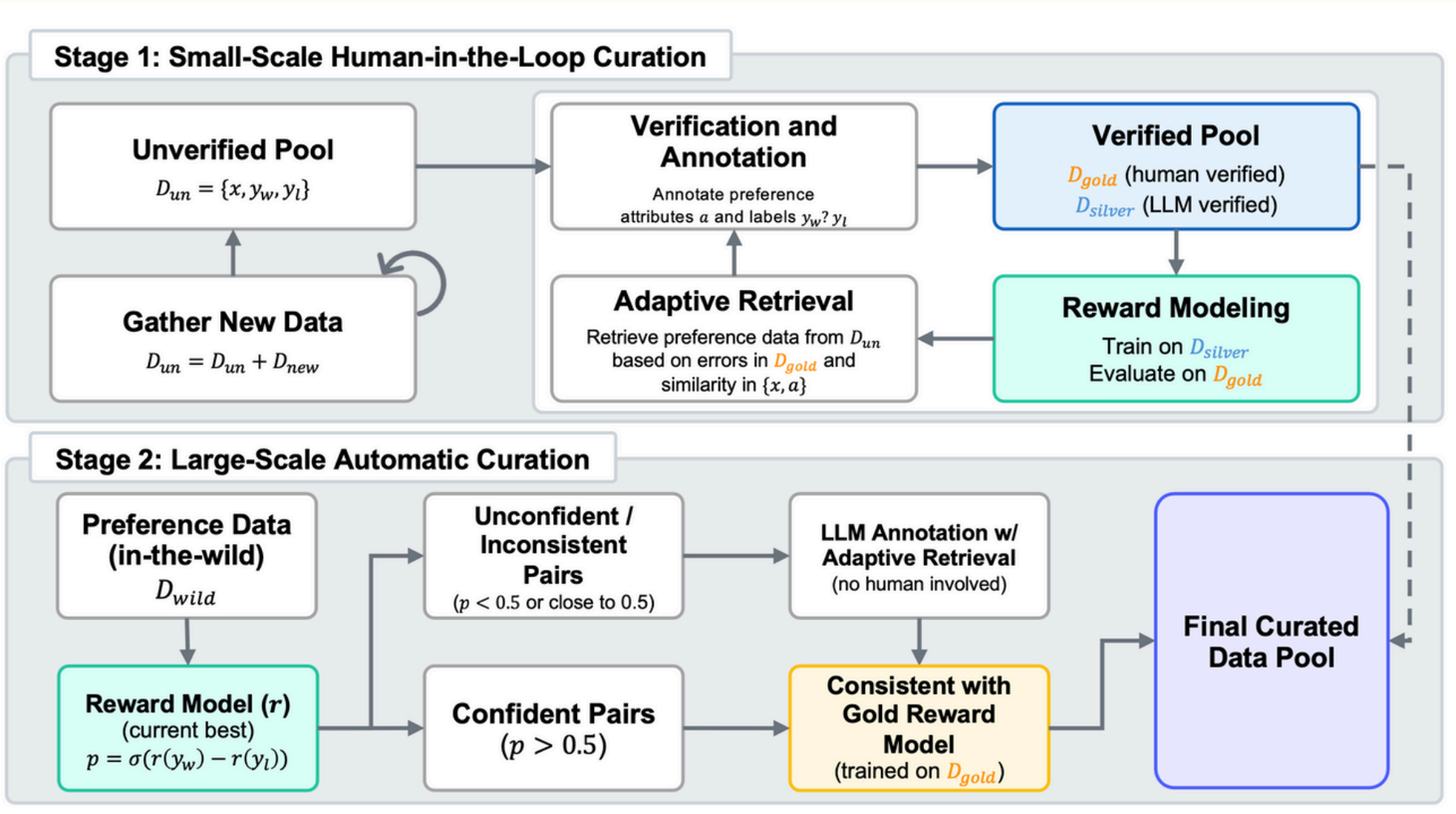
两阶段数据标注

阶段1：人参与的小规模标注

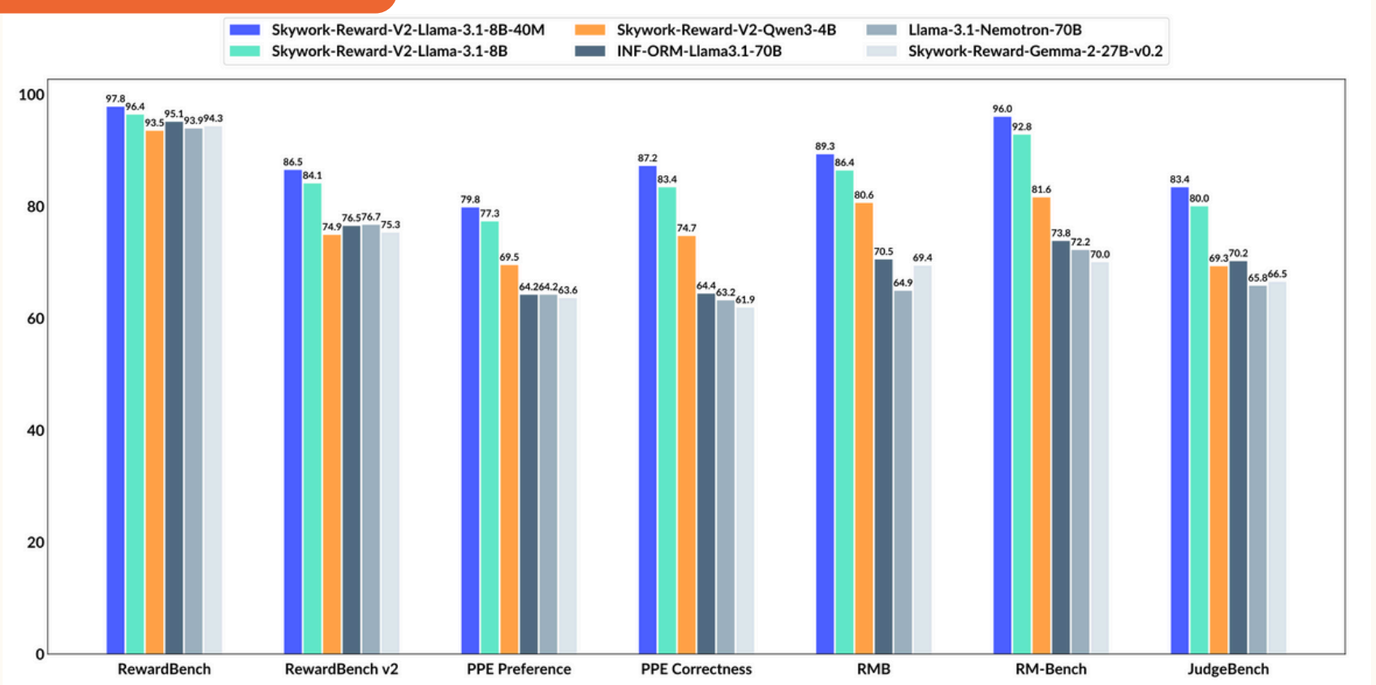
- 人工严格标注一小批高质量偏好数据，llm辅助标注一小批次高质量偏好数据。
- 用llm标注的数据训练reward model，用人标注的作为验证集。
- 利用reward model的预测错误数据，去从已有偏好数据中检索相似的，再让一组llm去二次标注。这一步相当于找更多reward model预测不对的偏好数据。
- 重复迭代以上过程，可以提升reward model和扩展更多高质量的偏好数据。

阶段2：大规模自动标注

- 借助训练好的reward model先对海量偏好数据过滤，再用llm二次标注



部分实验结果



思考

相比于其他llm厂家，在open reward model领域，Skywork似乎投入精力更多，可能背后有自己的原因吧，总之，能开源模型总是好的。