Learning to Reason without External Rewards

开源代码：HTTPS://GITHUB.COM/SUNBLAZE-UCB/INTUITOR

Xuandong Zhao*
UC Berkeley
xuandongzhao@berkeley.edu

Zhewei Kang*
UC Berkeley
waynekang@berkeley.edu

Aosong Feng
Yale University
aosong.feng@yale.edu

Sergey Levine
UC Berkeley
svlevine@berkeley.edu

Dawn Song
UC Berkeley
dawnsong@berkeley.edu

简介

本文设计了一种隐式的reward计算方式，即利用llm自身的置信度作为reward，其核心思想基于一个观察结果：llm在遇到**难题**时往往表现出较低的置信度。如何定义置信度呢？它等于llm对vocab的输出概率分布与均匀分布之间的平均 KL 散度

$$\textbf{Self-certainty}(o|q) := \frac{1}{|o|}\sum_{i=1}^{|o|}\mathrm{KL}(U \parallel p_{\pi_\theta}(\cdot|q,o_{<i})) = -\frac{1}{|o|\cdot|\mathcal{V}|}\sum_{i=1}^{|o|}\sum_{j=1}^{|\mathcal{V}|}\log\left(|\mathcal{V}|\cdot p_{\pi_\theta}(j|q,o_{<i})\right)$$

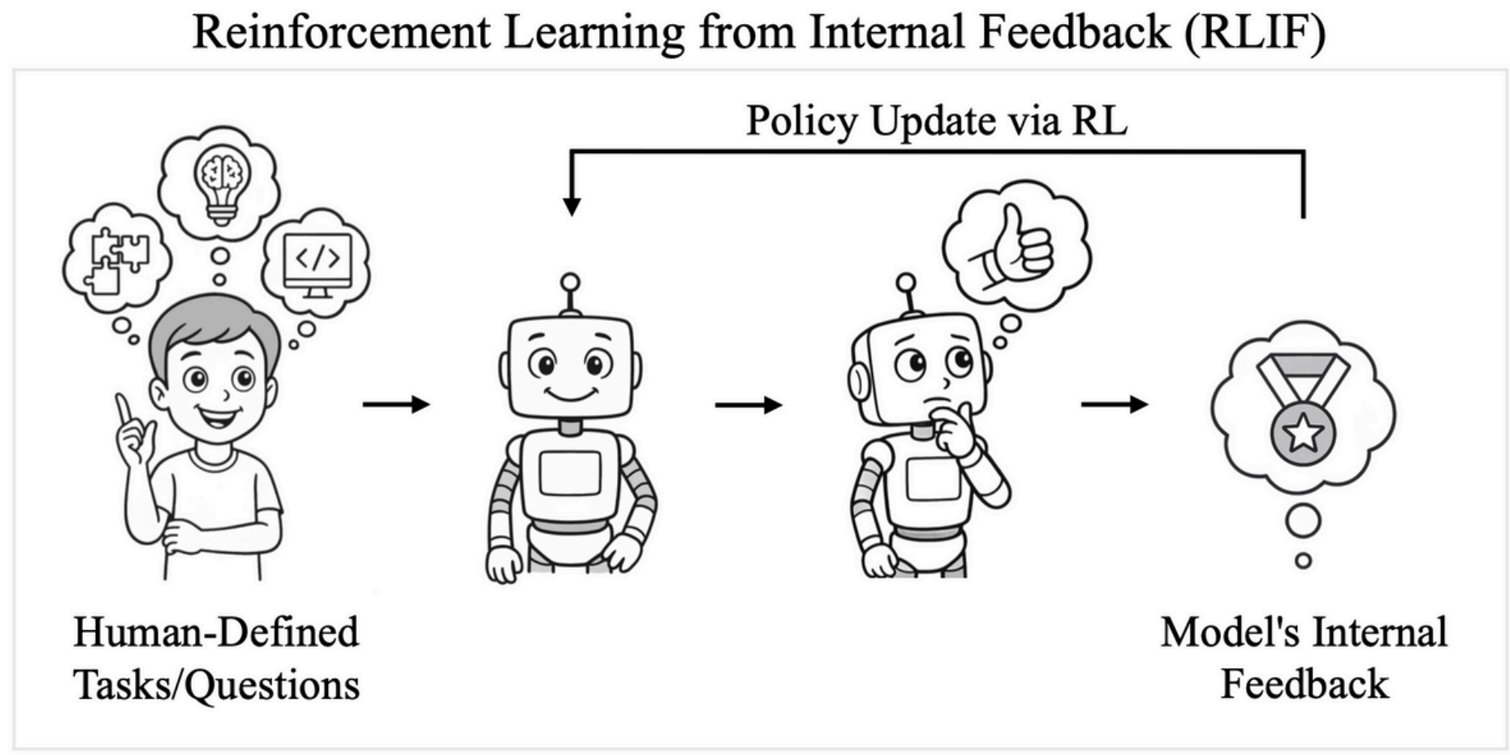
背景

不论是RLHF还是RLVR 都依赖明确的reward，前者依靠reward model，为了建模reward model需要标注偏好数据，后者依赖数据的ground truth（比如数学题的标准答案），但是很多领域难以找到适合RLVR的训练数据。
本论文延续隐式reward的思路：在没有标准答案和人类反馈的场景下，如何用RL训练llm？关键是如何定义“隐式reward”。本文作者设计了self-certainty作为reward。

实验设置

- 框架：open-r1
- 实验对象：Qwen-2.5 1.5B/3B/7B/14B和Llama3.2-3B-Instruct
- 强化学习算法：GRPO，保留kl loss
- reward function：self-certainty和KL两项

REWARD



部分实验结果

Table 1: Performance comparison of various methods on the GSM8K, MATH, LCB, CRUXEval-O, MMLU-Pro, and AlpacaEval benchmarks. The INTUITOR-Code variant is trained on Codeforces data with a smaller learning rate and fewer training steps. All evaluations are obtained with the chat inference template, except for MMLU-Pro.

Model	Training Data	GSM8K	MATH500	LCB	CRUX	MMLU-Pro	AlpacaEval
Qwen2.5-1.5B Results							
Base	-	0.002	0.090	0.000	0.000	0.297	2.10
+ GRPO	MATH	0.747	0.560	0.056	0.328	0.315	4.03
+ INTUITOR	MATH	0.711	0.530	0.099	0.296	0.310	4.28
Qwen2.5-3B Results							
Base	-	0.673	0.544	0.093	0.236	0.377	3.72
+ GRPO	MATH	0.826	0.636	0.085	0.341	0.403	6.91
+ GRPO-PV	MATH	0.820	0.636	0.086	0.299	0.398	6.17
+ INTUITOR	MATH	0.792	0.612	0.153	0.416	0.379	7.10
+ INTUITOR-Code	Codeforces	0.743	0.572	0.153	0.411	0.386	4.16

思考

我一直觉得隐式REWARD类型的工作挺有意思的，如果真的能通过无监督的方式让LLM提升能力，那可太棒了，此处省略一万个优点。
再来思考下可能存在的不足，将置信度作为REWARD，LLM可能对一些“幻觉”内容信心满满，这个时候还进行鼓励，岂不是对错误问题越陷越深？REWARD HACKING的风险是否也比较高，模型可能学会制造高置信度而非正确的答案，最终优化的也许只是“看起来自信”，而不是“真正有理有据”的推理。