arxiv.org/abs/2508.05004

R-Zero: Self-Evolving Reasoning LLM from Zero Data

R-Zero: 基于Challenger-Solver协作来提升llm math reasoning能力

Chengsong Huang[1,2]✉, **Wenhao Yu**[1]✉, **Xiaoyang Wang**[1], **Hongming Zhang**[1], **Zongxia Li**[1,3], **Ruosen Li**[1,4], **Jiaxin Huang**[2], **Haitao Mi**[1], **Dong Yu**[1]

[1]Tencent AI Seattle Lab, [2]Washington University in St. Louis,

[3]University of Maryland, College Park, [4]The University of Texas at Dallas

chengsong@wustl.edu; wenhaowyu@global.tencent.com

代码地址：https://github.com/Chengsong-Huang/R-Zero

简介

本文提出R-Zero，一种无需训练集，完全依赖llm自己合成数据训练自己(self-evolving)就能提升其math reasoning能力的训练方法。简单来说，R-Zero将实验对象llm初始化为两个独立的角色，负责生成数学问题的Challener和负责解题的Solver，二者均使用GRPO训练，相互配合迭代进步。

Challenger训练过程：重点是如何得到Challenger的训练数据和设计reward function。
Challenger生成一批数学题，用当前的Solver解答多次，统计每道题正确率。作者认为：一个好的Challenger应该生成对Solver来说正确率约50%的题目，这样Solver学习效率最高。为此设计了uncertainty reward，同时为了避免batch内有重复生成的题目，reward function中引入了重复惩罚，用GRPO训练。

Solver训练过程：这个比较简单，让Challenger生成一批数学题，由Solver解答多次投票得到每道题的伪标签，然后过滤掉太简单和太难的题目，剩下的就是训练集，然后用GRPO训练。

以上训练过程迭代多次。

背景

本文属于llm reasoning方向的工作，如何提升llm的reasoning能力，自从DeepSeek-R1之后，RLVR就成了学术界主流，但是RLVR需要有ground truth(基本是人写的)并且容易计算reward值的训练集，这在一定程度上限制了RLVR方法的扩展性。

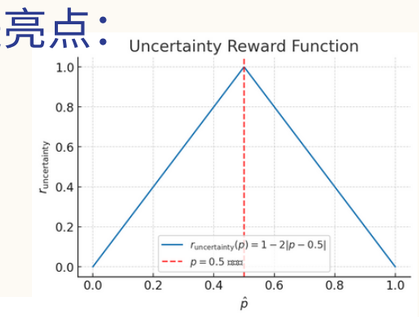
能否不依赖数据标注，或者说不依赖人工创建的训练集，让llm自己生成数据然后反过来训练自己来提升reasoning能力呢？本文提出的R-Zero就是一种方案。

实验设置

- 模型：两个qwen(Qwen3-4B-Base、Qwen3-8B-Base)和两个Llama(OctoThinker 3B/8B)
- 实验中生成的question都是数学领域问题
- 代码基于EasyR1 codebase
- Challenger的reward function是亮点：

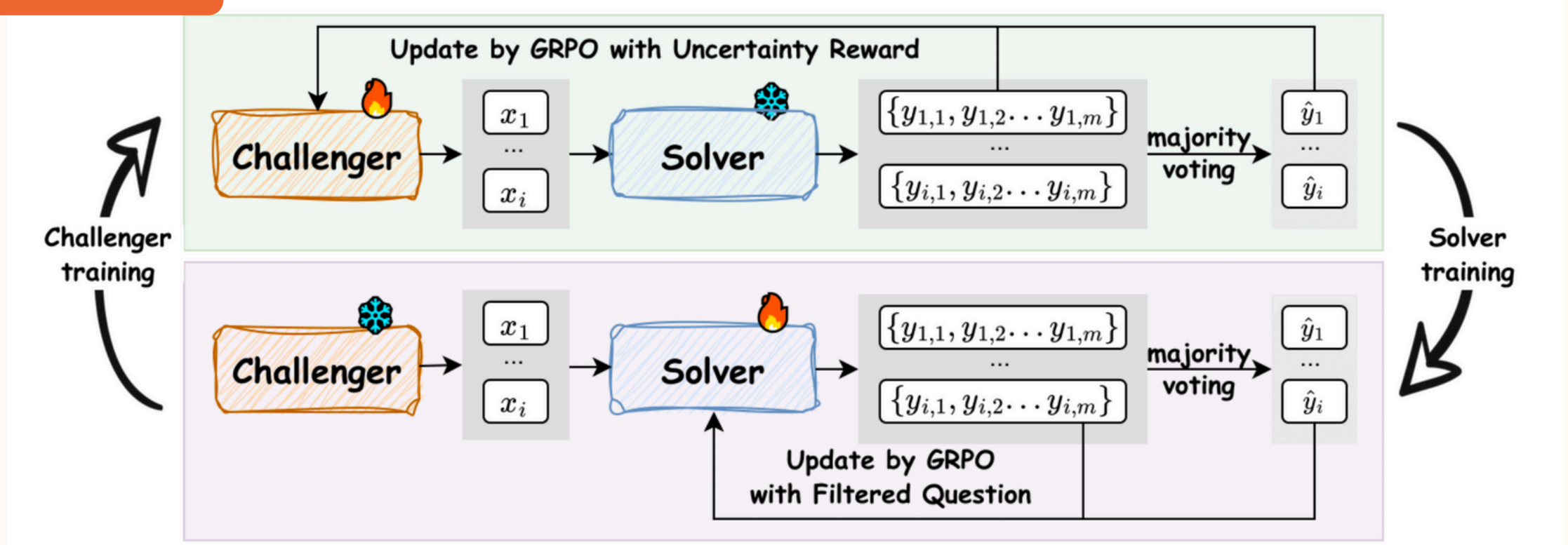
$$r_{\text{uncertainty}}(x;\phi) = 1 - 2\left|\hat{p}(x;S_\phi) - \frac{1}{2}\right|$$

$$r_{\text{rep}}(x_i) = \lambda \frac{|C_k|}{B}$$



$$r_i = \max\left(0, r_{\text{uncertainty}}(x_i;\phi) - r_{\text{rep}}(x_i)\right)$$

R-Zero框架



思考

当llm的能力已经超越人类，那个时候人类创建的训练集，已经不能提供对模型来说是新知识的时候，如何继续提升llm的能力呢？self-play或许是一个值得深入探索的方向，再就是可以结合之前的Absolute Zero论文一起看。