# Reinforcement Pre-Training

arxiv.org/abs/2506.08007


Qingxiu Dong*†‡ Li Dong*†

Yao Tang† Tianzhu Ye†§ Yutao Sun†§ Zhifang Sui‡ Furu Wei†◇

† Microsoft Research

‡ Peking University

§ Tsinghua University

https://aka.ms/GeneralAI


## 简介

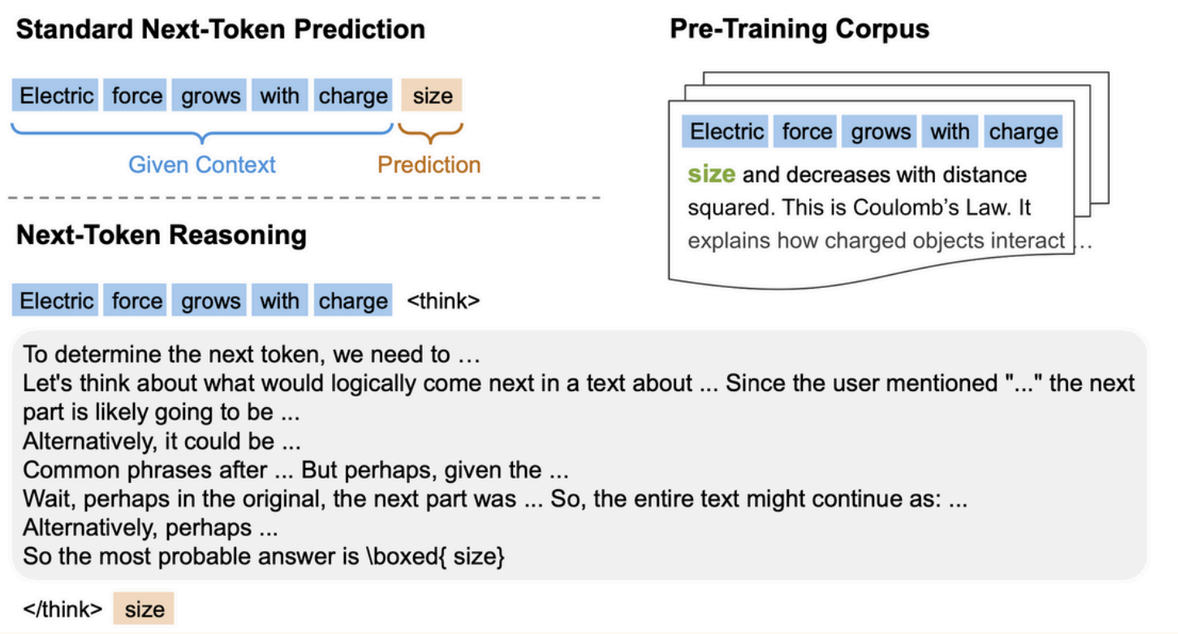

本文提出了一种名为Reinforcement Pre-Training（RPT）的新型预训练范式，将传统预训练阶段的next-token prediction任务重构为强化学习任务next-token reasoning，核心在于将每一步的token预测视为一个“推理决策过程”，并通过RLVR的方式来训练。具体而言，模型在预测下一个token之前，先进行思考再预测结果。这种将自监督学习（ssl）和RLVR结合的方式，为通用领域的可扩展RLVR提供了一种解决方案。


## 背景

LLM之所以具备强大的能力，核心在于预训练阶段采用scalable自监督学习范式：在超大规模文本语料上进行的next-token prediction训练任务。与此同时，RL也在后训练阶段发挥重要作用，包括RLHF和RLVR，但是现阶段的RLHF和RLVR都受制于数据的约束，前者依赖人工标注偏好数据训练reward model，后者需要的具有金标准的数据集也往往局限在数学、编程等少数领域，一句话，如何实现可扩展的RL (scalable RL)训练是个大问题，要想完全scalable，就不应该受制于人工标注数据。

我们可以思考一个问题：既然在预训练阶段可以用自监督学习ssl，那么能不能把ssl和RLVR结合起来呢？核心是如何设计一个合理高效的ssl任务，本文作者用的是next-token reasoning 任务。

## RPT



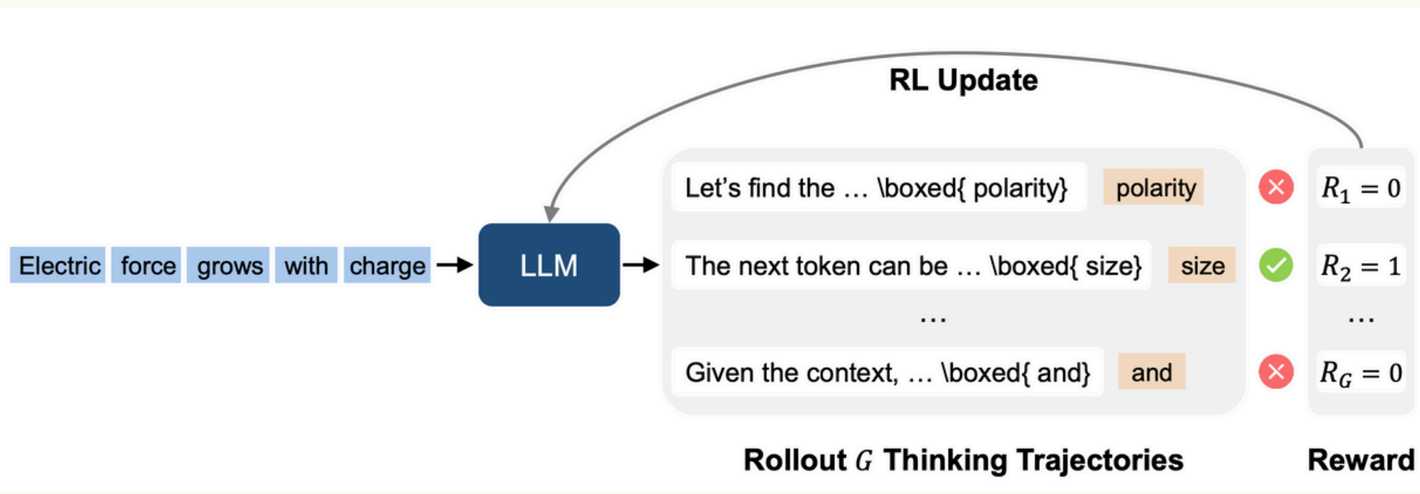

next-token reasoning和next-token prediction的区别是llm先思考再预测下一个token。

RL算法：GRPO

模型：Deepseek-R1-Distill-Qwen-14B

数据集：OmniMATH

reward：prefix matching reward




本文实验用的模型和数据集不太能说明可扩展性和通用性RLVR实验结果先不看了。相信作者后续肯定会在更大规模数据集做训练

## 思考

这几个月大量的RLVR工作出现，一下子把RLVR推到了llm前沿，只需要提供prompt和answer，就能让llm自由探索推理路径生成answer，确实太诱人了。一个自然地想法是，能不能把规模扩大，做大规模RLVR训练？问题出现了，上哪去找这么多有金标准answer的数据呢？数学和编程领域似乎有一些，但感觉也不多，我们想的大规模RLVR肯定是通用领域大规模RLVR，怎么做呢，一种很可行的方案是用自监督学习ssl，ssl的核心是设计一个自监督学习任务，本文作者巧妙地将预训练阶段的next-token prediction拿过来做ssl + RLVR，非常精彩。至于本文的实验则做的有点那个，用的是数学数据集，也不通用啊，并且reward也奇怪，直接比较next token是否正确不就行了？为什么设计prefix matching reward，我的猜测是因为训练集的answer是字符串，只能设计这样的reward做训练了，当然模型也不具有说服力，还有一点，毕竟训练RL的成本放在那里，现阶段用来做pre-training不敢想象，题目直接用“Reinforcement Pre-Training”似乎也有点不太合适。但是，以上都不是批评，用next-token prediction任务做ssl + rl for rl scaling 的思路一点毛病没有，非常好，好得很，我相信只有ssl(implicit reward)才是让rl做Scaling的**唯一**路径。当然后续是如何设计更加高效的ssl任务，希望看到更多精彩的论文。

@机器爱学习