1. Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models

Accession number: 20250359596

Authors: Zhang, Zizhuo (1); Zhu, Jianing (1); Ge, Xinmu (2, 3); Zhao, Zihua (2); Zhou, Zhanke (1); Li, Xuan (1); Feng, Xiao (1); Yao, Jiangchao (2); Han, Bo (1)
Author affiliation: (1) TMLR Group, Department of Computer Science, Hong Kong Baptist University, Hong Kong; (2) CMIC, Shanghai Jiao Tong University, China; (3) Shanghai Innovation Institute, China
Corresponding authors: Yao, Jiangchao(Sunarker@sjtu.edu.cn); Han, Bo(bhanml@comp.hkbu.edu.hk)

Source title: arXiv
Abbreviated source title: arXiv
Issue date: October 3, 2025
Publication year: 2025
Language: English
E-ISSN: 23318422
Document type: Preprint (PP)
Repository: arXiv

Abstract: While reinforcement learning with verifiable rewards (RLVR) is effective to improve the reasoning ability of large language models (LLMs), its reliance on human-annotated labels leads to the scaling up dilemma, especially for complex tasks. Recent self-rewarding methods investigate a label-free alternative to unlock the reasoning capabilities of LLMs, yet they frequently encounter the non-negligible training collapse issue, as the single-view supervision signal easily forms the self-consistent illusion, yielding the reward hacking. Inspired by the success of self-supervised learning, we propose Co-rewarding, a novel self-supervised RL framework that improves training stability by seeking complementary supervision from another views. Specifically, we instantiate Co-rewarding in two ways: (1) Co-rewarding-I is a data-side instantiation that derives reward signals from contrastive agreement across semantically analogous questions; and (2) Co-rewarding-II is a model-side instantiation that maintains a slowly-updated reference teacher with pseudo labels to realize self-distillation. Intuitively, such instantiations introduce different levels of discrepancy to increase the difficulty of training collapse on trivial reasoning solutions. Empirically, Co-rewarding exhibits stable training across various setups, and outperforms other self-rewarding baselines by +3.31% improvements on average on multiple mathematical reasoning benchmarks, especially by +7.49% on Llama-3.2-3B-Instruct. Notably, Co-rewarding reaches or even surpasses RLVR with ground-truth (GT) label in several cases, such as a Pass@1 of 94.01% on GSM8K with Qwen3-8B-Base remarkably higher than GT. Our code is publicly available at https://github.com/tmlr-group/Co-rewarding. © 2025, CC BY.

Number of references: 0
Main heading: Personnel training
Controlled terms: Computational linguistics - Reinforcement learning - Self-supervised learning - Supervised learning - Teaching
Uncontrolled terms: Complex task - Ground truth - Label free - Language model - Reasoning ability - Reasoning capabilities - Reinforcement learnings - RL framework - Scaling-up - Two ways
Classification code: 901.2 Education - 912.3 Personnel - 1101.2 Machine Learning - 1101.2.1 Deep Learning - 1102.1 Computer Theory, Includes Computational Logic, Automata Theory, Switching Theory, Programming Theory
Numerical data indexing: Percentage 3.31E+00%, Percentage 7.49E+00%, Percentage 9.401E+01%
DOI: 10.48550/arXiv.2508.00410
Compendex references: YES
Preprint ID: 2508.00410v2
Preprint source website: https://arxiv.org
Preprint ID type: ARXIV
Database: Compendex
Data Provider: Engineering Village