Week 10: Deep Dive into RLVR with nano-aha-moment

BEH Chuen Yang

Abstract

This report explores the use of RLVR for base model post-training in two parts. Firstly, we explore how RLVR is implemented in the nano-aha-moment repository, which applies the GRPO algorithm to train a Qwen model on a task called countdown. Secondly, we adapt the code to train on the GSM8K dataset, a grade school level math problem dataset. We modify the prompt template and the equation reward function to suit the GSM8K task. The results show that the adapted model can effectively solve grade school level math problems.

1 Introduction

Apparently continuing in the nanoGPT (**?**) tradition, the nano-aha-moment repository provides a minimalistic implementation of RLVR (Reinforcement Learning with Verifiable Rewards) in the R1-Zero fashion for base language model post-training.

References