NLP Project Errata

Matthew LeBar

Februray 1, 2026

There are two errors in my initial report on RLVR.
First, I incorrectly assert that training rewards are not representative of the model's training progress because the actions are drawn from the reference policy. This is incorrect; there are two loops of policy updating, and the actions are drawn from the policy in the outer loop, which therefore updates less frequently but is distinct from the reference model (which is updated even less frequently). It is true that the training rewards are not indicative of model quality, but that is in fact because rewards are always averaged over a group of responses.
Second, on page 3 I say the advantage function of action a at state s is given by $Q(s, a) - v(a)$. This of course should be $Q(s, a) - v(s)$.