



BetterEnv World

scalingX

RsiingSI · Independent Researcher

20261017

When we build an agent environment, how should we evaluate the quality of the environment itself?

We propose BetterEnv World: before post-training, give the agent environment a health check first.

When training agents, we keep tuning the model, the reward, and the RL algorithm—yet we rarely evaluate the environment itself.

But environment quality directly determines whether post-training makes the model stronger, makes it weaker, or teaches it to hack the sandbox.

A bad environment carries three kinds of poison:

1. **No signal:** every reward in the group is identical, the GRPO advantage is 0, the gradient is 0—wasted training.
2. **Fake signal:** the reward correlates only weakly with true success; the more you train, the further it drifts.
3. **Hackable:** the model discovers that gaming format, length, or tool-call counts earns score, and learns to hack the sandbox.

We propose BetterEnv World: a quality-diagnosis protocol for agent environments, plus a multi-stage multi-agent debug pipeline.

Core ideas:

- **Multi-stage multi-agent debug:** designer / solver / verifier / attacker / fixer automatically surface task ambiguity, reward loopholes, tool instability, and reward hacking.
- **Fast GRPO proxy with a tiny model:** look beyond loss—at within-group reward variance, the fraction of non-zero advantages, and the gradient signal-to-noise ratio.
- **Controlled bias injection:** on hard tasks a small model often fails everything, so within-group advantage is 0. We inject a small-coefficient, switchable, preferably potential-based bias as a diagnostic probe, to check whether the model learns real capability or merely games the proxy metric once a signal appears.
- **The final criterion is not whether the loss moved, but:** does held-out true success improve? Is the model more robust under adversarial attack? Is the result reproducible across models and seeds? Can small-model diagnostics predict large-model post-training gains?

We want to answer one question:

Can we tell, before training a large model, whether an agent environment is worth training on—and whether it will ruin the model?

More detail soon. Discussion and collaboration welcome.