HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA




RESEARCH · 2026

# Reflective Prompted Policy Optimization (R2PO)

*Trajectory-Grounded Revision and Salience Bias*

Rahaf Abu Hara · Vaibbhav Murarri · Claudio Zito
Laboratory of AI and Robotics Research — Heriot-Watt University, Dubai

# The Problem

*Optimizing an RL agent's policy from reward signals alone*

What actually happened?

- 🕳️ Fell in a hole on step 3?
- 🔁 Stuck in a loop?
- ✅ 19/20 succeeded—one catastrophic failure?

*Different behaviors… same final score.*

> *Policy optimization becomes blind trial-and-error.*

The optimizer only sees:

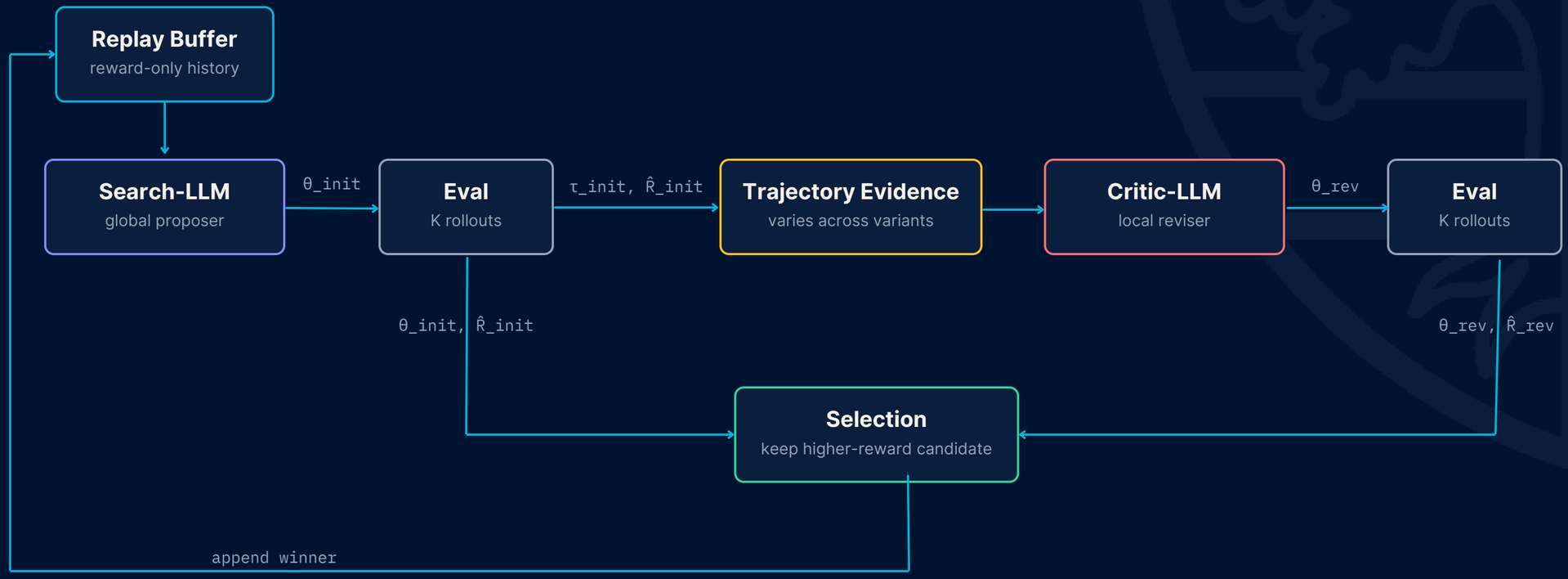



- × No behavioral evidence
- × No explanation
- × No diagnostics

LAIRR
LABORATORY OF AI AND ROBOTICS RESEARCH

# R2PO Framework

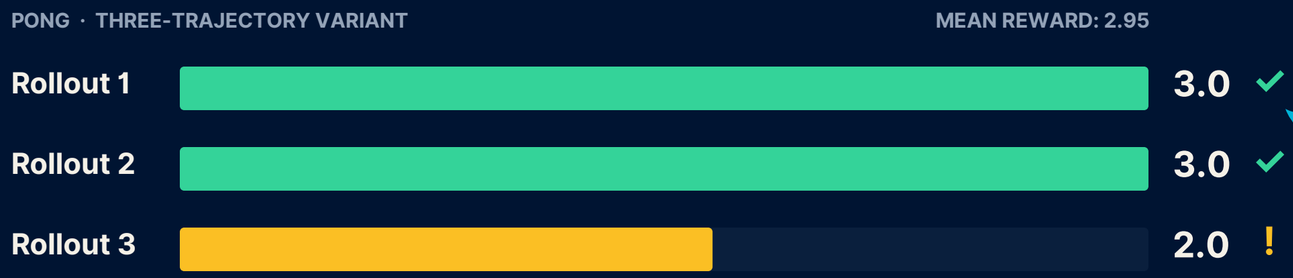
*Two LLMs. Two jobs.*




*Trajectories as first-class evidence.*

# Salience Bias

*When the Critic-LLM overweights worst-case rollouts*




**What the Critic-LLM focused on:**

*"From the worst rollout we saw that the policy never moved..."*

rewrote 6 parameters → reward dropped to **0.15**

## 76.6% of regressions on CartPole

**SALIENCE BIAS**

*When shown multiple rollouts of a policy, the Critic-LLM systematically overweights the worst (most striking) trajectory in its diagnosis and edits, even when most rollouts indicate good performance.*

# The Fix: Three Components

*How R2PO mitigates salience bias*

**1**

## Median Trajectory

*typical, not worst-case*

**2**

## Aggregate Statistics

*"19/20 ok" > one trace*

**3**

## Revision Rule

*if it works, leave it alone*

Acceptance: 70% ↠ **87%**

Regressions: **cut by half**

> *Representative evidence beats vivid evidence.*

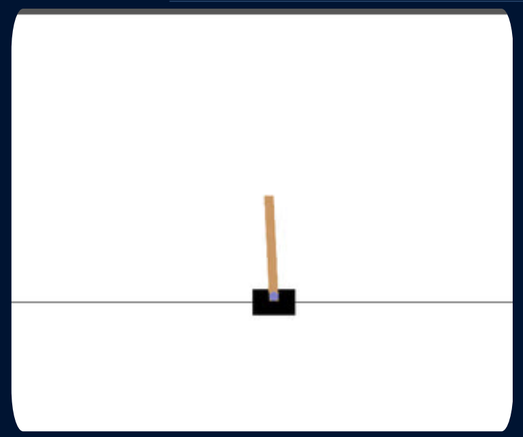
# Results and Performance

*R2PO across ten environments with a 20B open-weight model*

**10 / 10**
environments
won against deep RL and prior LLM-based baselines.

**20B**
open-weight
*(not GPT-4o)*

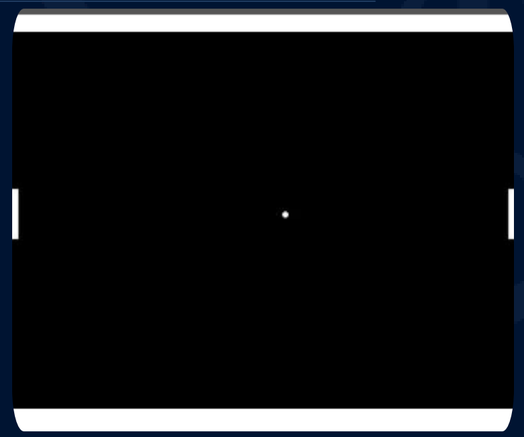
**500**
episodes to
near-optimal CartPole
*(vs. 4,000 baseline)*



**CartPole**



**FrozenLake**



**Pong**

*No fine-tuning · No gradient updates · Just better feedback*

LAIRR
LABORATORY OF AI AND ROBOTICS RESEARCH

# Trajectories aren't background.

## *They're the lesson.*

| | |
|---|---|
| **Code** | github.com/HWUDLabAIRoboticsResearch/R2PO |
| **Paper** | arxiv.org/abs/2605.08315 |
| **Website** | hwudlabairoboticsresearch.github.io/R2PO |
| **Contact** | rha4001@hw.ac.uk · C.Zito@hw.ac.uk · vm81@hw.ac.uk |

Website

Funded by Dubai RDI Grant — CogWaters Project (ID: 2025/DRDI0163)