AIMS WORKSHOP / COLM 2026

大阪大学
THE UNIVERSITY OF OSAKA

Replayable Trajectory Dataflow

Auditing Reward Measurement under Optimization and Drift in RLVR

Zepeng Zhang The University of Osaka

Preserve the evidence behind training rewards.
Re-evaluate historical outputs, trace disputed signals, and plan where to restart.

Why retain trajectories?

In reinforcement learning with verifiable rewards (RLVR), the verifier is a measurement instrument *inside* the training loop.

Strategic optimization

The policy exploits a flawed proxy. Training reward rises while the intended quantity falls.

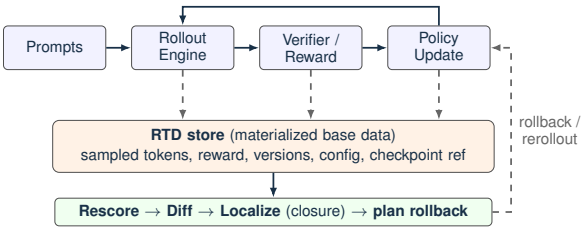
Drift

A verifier or reward function silently changes mid-run. The loop now optimizes a different reading.

- Checkpoints preserve state. Aggregate metrics can hide failures.
- Re-sampling a checkpoint need not reproduce the outputs that were scored.

Neither alone identifies which historical measurements shaped an update.

One store, three audit queries



Rescore
Apply a pinned reference to the same retained outputs.

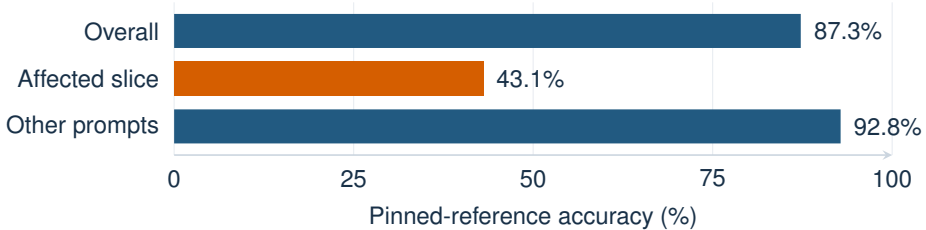
Diff
Compare reference readings with the signals actually consumed.

Localize
Follow consumption and state dependencies to affected updates and checkpoints.

Keep what the audit cannot reconstruct; derive the rest only with pinned dependencies. No query re-runs the rollout engine.

Aggregate accuracy can hide a failure

13.7% of prompts form a degraded slice.



Controlled arithmetic regression, final 50 updates. All 5,483 mismatches lie in the injected slice; this checks the mechanism, not discovery of an unknown hack.

Complete trajectory-level evidence exposes a subgroup that the aggregate masks.

A real-model training run, audited offline

41,600
retained trajectories

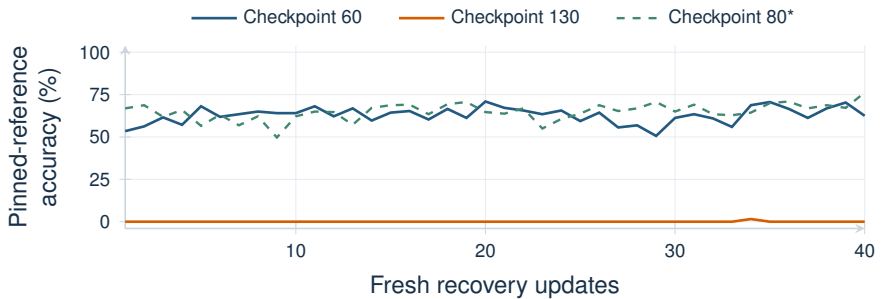
2.99 s
CPU query stages;
no new sampling

VeRL / Qwen2.5-Math-1.5B on MATH, injected verifier regression.



First disagreement: update 61. Rollback boundary: checkpoint 60.

Recovery from the boundary



After 40 fresh updates: **62.5%** from checkpoint 60 vs. **0%** from 130. *Later checkpoint-80 arm: **76.25%**, with a partial independent audit. Single-arm training-batch scores, not held-out accuracy.

This boundary need not give the best recovery. Resume with fresh rollouts under the repaired verifier.

Guarantees, limits, and links

Closure

With complete lineage, closure identifies descendants of disputed signals and a saved rollback boundary outside those dependencies.

Scope of the result

- A reference is required; it can share the original verifier's blind spots.
- Affected ancestry does not imply poor recoverability.
- One injected incident on a real-model stack; 91.49 MB sidecar. The 2.99 s measurement excludes startup and store opening.
- Stale trajectories are evidence, not training data.

github.com/PeterWrighten/rtd

colm.cc/virtual/2026/3454

zhang.zepeng@ist.osaka-u.ac.jp

Audit evidence supports recovery decisions; ancestry alone does not select the best repair.