

SAMPLE DELIVERABLE · SYNTHETIC DATA · MONTHLY VERDICT

Harness Watch verdict

Support-ticket triage agent, the same synthetic workload as the sample Harness Scorecard, one month after the candidate model was adopted behind the gate. A new version of that model was released this month. Its scorecard was rerun on the same 50 cases and gate as the last accepted configuration.

Hold

VERDICT
Quality and cost improve, but two reply-template cases now fail the gate.
Next check: the next model release. Next quarterly review: January 12, 2027.

SCORECARD CHANGE AGAINST THE LAST ACCEPTED CONFIGURATION

Measure	Last accepted	New release	Change
Quality	50 cases · 51 attempts · 45 accepted (88%)	50 cases · 50 attempts · 47 accepted (94%)	Better on long tickets
Speed	Median 7.1 s	Median 7.3 s	Unchanged within noise
Cost	Index 84 per accepted task	Index 76	Lower
Control	0 unauthorized actions · 0 failed checks	0 unauthorized actions · 0 failed checks	Unchanged since the refund fix
Drift	Baseline	2 gate failures on the reply template	Blocks adoption

WHAT CHANGES THIS MONTH

Next steps

- Fix the reply-template formatter for the two failing cases.
- Re-run the gate on the fixed configuration.
- Adopt behind the gate if all cases pass.

Scaffolding to delete on adoption

- The retry wrapper around the ticket classifier.
- The long-ticket truncation step.

Autonomy

Ticket routing holds at observe-after. Refunds hold at review-before. Both are revisited at the quarterly review with the month's intervention data.

Scope: a synthetic workload built to show the format of a Harness Watch verdict. The agent, the release and the results are illustrative and do not describe a client system or a specific provider's model. It continues the workload in the sample Harness Scorecard. Cost is shown as an index so the format carries no prices. A client verdict runs on the client's own cases and gate.