Dual-Path Anti-Drift Reward Architecture for Reliable Long-Horizon LLM Agents

A Mechanistic Simulation Study with Formal Analysis

Yuval Fradkin

September 2026

Abstract

Long-horizon language model agents face a structural pressure toward *false success*: as task complexity grows, the expected reward from a confident but unverified answer can exceed the expected reward from continued evidence gathering. This paper proposes **Dual-Path Anti-Drift** (DPAD), an architecture that structurally separates the solver from its quality assessor. A *primary path* pursues solutions under a reward function that favours verified correctness; an *independent process critic*, isolated in context and reward, searches for rewarding shortcuts and unsupported confidence escalations. Confirmed critic findings trigger reward deductions on the primary path; evidenced stopping earns partial credit. A deterministic coordinator enforces state transitions; an append-only evidence ledger prevents silent revision.

We derive four formal results for the three-action softmax model. **P1** (monotone suppression): the shortcut penalty strictly reduces shortcut-selection probability at every step. **P2** (per-step hazard amplification): higher error coupling strictly increases the conditional per-step shortcut-derived false-success hazard through two simultaneous channels. **T1** (structural FSR bound): the analytical supremum of the utility gap is $\Delta_{\max} = 1.822$, derived from the simulator's utility equations; at default fixed-review parameters ($\rho = 0.30$, $\lambda = 2.2$), $\lambda\hat{q}(0.30, 0.72) = 0.99 < \Delta_{\max}$ and the bound is trivially 1.0; two distinct thresholds govern usefulness: pointwise penalty dominance requires $\lambda > 4.05$; a bound below 1.0 requires $\lambda > 6.37$; the bound's value is structural, showing how FSR_{sc} would decrease monotonically (and asymptotically exponentially) with $\lambda\hat{q}(\rho, 0.72)$. **P3** (Bayes-optimal critic threshold): under an independent verifier, the threshold policy at $p^* = b/(a + b + c)$ weakly dominates constant reporting strategies. These results hold for the formal model, not for trained LLMs. Throughout this paper, *hazard* denotes the probability of an incorrect success claim within the formal model; it does not denote expected real-world harm.

We evaluate DPAD via a mechanistic simulation across **384,000 episodes** (12 seeds, seeds 1101–1112). Key findings: an isolated critic reduces FSR from 69.68% to 42.79% and raises CCR from 29.22% to 47.93% (RCI 0.336 vs. 0.172, $p_{\text{Bonf}} < 0.005$); full fixed-review achieves the lowest FSR (37.61%) at lower CCR (40.68%), JAR 15.70%, and the second highest normalised cost (NC 4.25 vs. critic-only's 4.66); anti-drift reward alone is not significant on FSR ($p_{\text{Bonf}} = 0.051$) or CCR ($p_{\text{raw}} = 0.257$); no configuration Pareto-dominates the others across FSR, CCR, JAR, and cost; error coupling is the dominant moderator (FSR rises from 43% to 60% as ρ increases from 0 to 0.90).

Two simulation design features limit interpretation: the critic is invoked only after a shortcut is selected, not on all steps; and error coupling is modelled as linear damping of the critic's true-positive rate rather than as an empirically measured correlation. The critic reward R_C is not implemented, so P3 is not tested empirically. All results characterise the mechanistic model. Required LLM experiments are enumerated in Section 17.

Contents

1 Introduction

1.1 The Long-Horizon Reliability Problem

Language model agents deployed on multi-step tasks—security research, scientific literature synthesis, code auditing, legal review—face a structural incentive misalignment. Over many steps, a model trained to appear helpful accumulates pressure toward *false success*: the expected reward from a confident, unverified answer rises relative to the expected reward from continued evidence gathering. Two failure modes compound: *proxy reward capture* (optimising surface coherence rather than correctness) and *assumption drift* (early scope choices becoming load-bearing without revisitation, so the agent solves a shifted problem).

1.2 Why Structural Separation Matters

A critic sharing reward targets with the solver has reduced incentive to challenge shortcuts; joint optimisation tends to produce a critic that ratifies rather than audits. Process reward models [Lightman et al., 2024] provide step-level feedback but are typically trained alongside the policy they assess. DPAD places the critic on a separate path with a separate reward function and prohibits it from contributing to solutions, reducing the direct shared-objective incentive to ratify. Related empirical support comes from adversarial debate [Kenton et al., 2026]: critics under separate objectives reduced reward hacking in an RLAIF setting, though critics can also learn to satisfy the judge—the "flaw hunger" failure mode addressed by the false-positive penalty $-b \cdot \mathrm{FP}$ in R_C.

1.3 Relationship to Prior Work

AgentForesight [Zhang et al., 2026] deploys a runtime auditor on trajectories; Debate Training [Kenton et al., 2026] uses an adversarial critic; Agent-RRM [Fan et al., 2026] provides multi-dimensional trajectory feedback; iStar [Liu et al., 2026] allocates intermediate rewards. These works focus on intervention timing or reward density. DPAD's distinguishing combination is: a critic structurally prohibited from solution roles; symmetric rewards for verified solutions and justified non-solutions; an error-coupling deployment gate; and dynamic risk-sensitive review intensity, all supported by the formal analysis in Section 8.

1.4 Contributions

1. Formal definitions of false success, process drift, justified abstention, and error coupling (Section 4).
2. A dual-path architecture with isolated contexts, separate reward functions, a deterministic coordinator, and an append-only evidence ledger (Section 5).
3. Three formal results and one Bayes-optimal threshold result, each with complete proof (Section 8).
4. A reproducible mechanistic simulation across 384,000 episodes with component ablations and sensitivity analysis (Sections 11–13).
5. Transparent disclosure of five null or negative findings (Section 15).

2 Terminology Conventions

To avoid ambiguity between the proposed architecture and the simulator, the following terms are used consistently throughout:

Term	Meaning
DPAD architecture	The full proposed system including evidence ledger $\mathcal{L}$, reward functions R_P and R_C, the deterministic coordinator, and trained policies π_P and π_C. Not yet implemented on any LLM.
DPAD-Fixed-Sim	The `full_fixed` simulator configuration: critic enabled, anti-drift reward active, fixed review intensity, no trained policies.
DPAD-Adaptive-Sim	The `full_adaptive` simulator configuration: same as DPAD-Fixed-Sim but with adaptive review intensity.
Anti-drift mechanism	The overall architectural intervention that discourages shortcut selection.
Non-drift reward (R_{ND})	The positive reward component in R_P awarded for staying within δ of G_0. Implemented in the simulator by modifying the abstention utility.
Hazard (h_t)	Per-step probability of an incorrect success claim in the formal model: $h_t = P_t(S)\, w_t\, [1 - d_t(\rho)]$. Does not measure real-world harm severity.
Error-coupling parameter (ρ)	A simulator input that linearly reduces both the primary path's anticipated detection rate and the critic's actual TPR. Not identical to the empirical Pearson correlation between path failure indicators.

3 Related Work

RLHF and reward hacking. RLHF [Christiano et al., 2017, Ouyang et al., 2022] trains a reward model from human preferences and fine-tunes the policy; the reward model can be gamed [Gao et al., 2023]. Krakovna et al. [2020] catalogue specification gaming; Pan et al. [2022] show reward magnitude amplifies it.

Process supervision. Lightman et al. [2024] find process supervision superior to outcome supervision in their mathematical-reasoning setting. Uesato et al. [2022] compare process and outcome supervision. DPAD extends process supervision by placing the assessor on a structurally separate path.

Verifier and critic models. Cobbe et al. [2021] show that a separately trained verifier improves final-answer accuracy by ranking sampled candidate solutions. Li et al. [2023] combine chain-of-thought with diversity-guided verification. DPAD's critic is prohibited from the generator role.

Multi-agent debate. Du et al. [2023] improve factual accuracy through multi-model debate. Liang et al. [2023] encourage divergent reasoning. Kenton et al. [2026] provide related empirical support for separating critic and generator incentives: adversarial debate reduces reward hacking in their RLAIF setting, and critics can learn to satisfy the judge rather than assess quality—the structural analogue of flaw hunger addressed by $-b \cdot \mathrm{FP}$ in R_C.

Runtime auditing and trajectory feedback. Zhang et al. [2026] monitor trajectories and intervene at failure points. Fan et al. [2026] provide multi-dimensional reward signals. Liu et al. [2026] allocate intermediate rewards. DPAD complements these works with structural separation and formal analysis of the softmax action model.

Abstention and uncertainty. Kadavath et al. [2022] provide evidence that language models can estimate aspects of their own uncertainty under specific evaluation settings. DPAD adds a structural incentive: justified abstention earns partial reward; lazy abstention is penalised.

4 Terminology and Problem Formulation

Definition 1 (False success)**.** *The primary path declares an episode a success and the outcome is incorrect. In the simulator: an episode terminating via the shortcut action with an incorrect ground-truth result. Shortcut-derived correct outcomes are classified as* correct but unverified *and excluded from the FSR numerator.*

Definition 2 (Rewarding shortcut)**.** *An action yielding high local utility without advancing verified evidence toward G_0.*

Definition 3 (Process drift)**.** *$D_t = \text{dist}(G_t, G_0) > \delta$, where G_t is the effective goal at step t. In the simulator drift is approximated by shortcut selection. A fully operational definition for trained LLMs is left to future work.*

Definition 4 (Justified abstention)**.** *The episode terminates without a success claim, supported by evidence below threshold at or after minimum effort, or by horizon exhaustion.*

Definition 5 (Premature abstention)**.** *The episode terminates without a success claim with no evidential support.*

Definition 6 (Error-coupling parameter ρ)**.** *A scalar $\rho \in [0, 1]$ governing two distinct simulator quantities: (i) the primary path's* anticipated *detection rate $\hat{q}(\rho, I_t) = (0.70 - 0.25\rho)\, I_t$, which enters the utility calculation and shapes action selection; and (ii) the critic's* actual *true-positive rate $\text{TPR}(\rho) = \text{clip}(0.90 - 0.42\rho,\ 0.35,\ 0.92)$, which determines post-selection detection probability. Higher ρ depresses both simultaneously. This is a parameterised proxy; the empirical Pearson correlation between path failure indicators is not reported.*

State machine. The coordinator maintains $s \in \{\texttt{PASS}, \texttt{WARN}, \texttt{BLOCK}, \texttt{UNKNOWN}\}$. Only the coordinator may write s.

5 Proposed Architecture

5.1 Primary Path and Independent Critic

The primary path π_P accepts task specification G_0, tool set $\mathcal{T}$, and a read-only view of $\mathcal{L}$. It emits $e_P \in \{\texttt{STEP}, \texttt{SOLVED}, \texttt{UNKNOWN}, \texttt{ABORT}\}$ and cannot access the critic's context.

The critic π_C receives a sanitised summary of primary-path actions and emits $\{\texttt{CLEAN}, \texttt{FLAW}(f_1, \ldots, f_k)\}$. It identifies: (i) shortcuts, (ii) goal substitution, (iii) unsupported confidence escalations, (iv) omitted caveats, (v) evidence gaps. It *cannot* emit `SOLVED`. In the simulator the critic is invoked only after a shortcut action is selected—a design simplification that is a threat to validity (Section 16).

5.2 Deterministic Coordinator

Algorithm 1 Coordinator Transition Logic

Require: v_C, e_P, $\mathcal{L}$

if $v_C = \texttt{CLEAN}$ **and** $e_P = \texttt{SOLVED}$ **and** evidence $\geq$ threshold **then**
  $s \leftarrow \texttt{PASS}$
else if $v_C = \texttt{FLAW}(\cdot)$ **and** $e_P = \texttt{SOLVED}$ **then**
  $s \leftarrow \texttt{BLOCK}$; append flaw to $\mathcal{L}$
else if $v_C = \texttt{FLAW}(\cdot)$ **and** $e_P \neq \texttt{SOLVED}$ **then**
  $s \leftarrow \texttt{WARN}$; append flaw to $\mathcal{L}$
else if $e_P \in \{\texttt{UNKNOWN}, \texttt{ABORT}\}$ **and** $\exists$ uncertainty entry in $\mathcal{L}$ **then**
  $s \leftarrow \texttt{UNKNOWN}$ (justified)
else
  $s \leftarrow \texttt{UNKNOWN}$ (unjustified)
end if

5.3 Context Isolation and Evidence Ledger

Paths share no prompt, memory, or intermediate reasoning. $\mathcal{L}$ is append-only with entry status $\in \{\texttt{LIVE}, \texttt{SUPERSEDED}, \texttt{REFUTED}\}$; only the coordinator may set `REFUTED`. The evidence ledger is specified as a core architectural component but is not implemented in the current simulator (Appendix D).

6 Reward Design

6.1 Primary-Path Reward

$$R_P = R_{\text{verified}} + R_{\text{non-drift}} - \lambda P_{\text{shortcut}} - \eta P_{\text{unsupported}} + \alpha R_{\text{verified}} \cdot \mathbf{1}[\text{justified abstain}] - \beta R_{\text{verified}} \cdot \mathbf{1}[\text{lazy abstain}] \tag{1}$$

In the simulator, $\lambda = 2.2$ and the non-drift reward weight $R_{\text{ND}} = 1.0$ enter the softmax utilities directly. The shortcut utility is $u_s - \lambda \hat{q}(\rho, I_t)$, where the anticipated detection rate $\hat{q}(\rho, I_t) = (0.70 - 0.25\rho)I_t$ shapes action selection; and the abstention utility is increased by $R_{\text{ND}} \times (0.95$ if justified, else $0.20)$ when the anti-drift mechanism is active. The terms η, α, β, and R_C are part of the proposed DPAD architecture but are not computed in the simulator.

6.2 Critic Reward

$$R_C = a \cdot \text{TP} - b \cdot \text{FP} - c \cdot \text{FN} - d \cdot S \tag{2}$$

S penalises role collapse; $-b \cdot \text{FP}$ prevents flaw hunger. R_C is not computed in the simulator, so Proposition 3 (Section 8.4) is not tested empirically.

6.3 Influence of R_P at Inference

At frozen inference time, R_P can influence π_P via reranking, search-priority adjustment, retry triggering, or offline fine-tuning. The simulator implements reranking implicitly through utility modification. A planned LLM experiment must specify the mechanism before measuring R_P effects.

7 Dynamic Review-Intensity Controller

$$I_t = \begin{cases} 0.72 & \text{(fixed review)} \\ \text{clip}(0.25 + 0.75\, r_t,\ 0,\ 1) & \text{(adaptive review)} \end{cases} \tag{3}$$

where $r_t = \text{clip}(0.32d + 0.26q + 0.22s + 0.20p_t,\ 0,\ 1)$ combines difficulty d, dilemma density q, stakes s, and failure pressure p_t. User satisfaction cannot reduce I_t below a domain-mandated floor. The current adaptive parameterisation raises r_t in response to past failures rather than anticipating upcoming shortcuts; this is the cause of the calibration failure in Section 15.

8 Formal Analysis

All results apply to the three-action softmax model with temperature $T > 0$ and anticipated detection rate $\hat{q}(\rho, I_t) = (0.70 - 0.25\rho)I_t > 0$. They hold for the formal model; transfer to trained LLMs requires the experiments in Section 17.

8.1 Monotone Shortcut Suppression

Proposition 1 (Monotone shortcut suppression)**.** *For any fixed $\rho \in [0,1]$ and $I_t \in (0,1]$, let $\hat{q} = (0.70 - 0.25\rho)I_t > 0$ and $\tilde{u}_s = u_s - \lambda\hat{q}$. Then:*

$$\frac{\partial P(S)}{\partial \lambda} = -\frac{\hat{q}}{T} P(S)\,[1 - P(S)] < 0.$$

Increasing λ strictly reduces shortcut-selection probability at every step, for any fixed ρ and I_t.

Proof. The three-action softmax probability is: $P(S) = e^{\tilde{u}_s/T}/(e^{\tilde{u}_s/T} + e^{u_g/T} + e^{u_a/T})$. Since $\tilde{u}_s = u_s - \lambda\hat{q}$ and $\hat{q}$ does not depend on λ, we have $\partial\tilde{u}_s/\partial\lambda = -\hat{q}$. The softmax Jacobian gives $\partial P(S)/\partial\tilde{u}_s = P(S)(1 - P(S))/T$, so by the chain rule:

$$\frac{\partial P(S)}{\partial \lambda} = \frac{P(S)(1 - P(S))}{T} \cdot (-\hat{q}) = -\frac{\hat{q}}{T} P(S)[1 - P(S)] < 0. \qquad \square$$

$\square$

8.2 Per-Step Error-Coupling Amplification

Proposition 2 (Per-step hazard amplification)**.** *Define the per-step false-success hazard $h_t(\rho) = P_t(S)\, w_t\, [1 - d_t(\rho)]$, where $w_t = P(\text{shortcut incorrect})$ and $d_t(\rho) = \text{TPR}(\rho) \cdot I_t$. Then $\partial h_t/\partial\rho > 0$ at every step where $P_t(S) \in (0,1)$, $I_t > 0$, $\lambda > 0$, $w_t > 0$, and $d_t(\rho) < 1$.*

Proof. The penalised shortcut utility is $\tilde{u}_{s,t} = u_s - \lambda(0.70 - 0.25\rho)I_t$. Differentiating with respect to ρ:

$$\frac{\partial \tilde{u}_{s,t}}{\partial \rho} = -\lambda \cdot (-0.25) \cdot I_t = 0.25\lambda I_t > 0.$$

By the softmax Jacobian and chain rule:

$$\frac{\partial P_t(S)}{\partial \rho} = \frac{P_t(S)[1 - P_t(S)]}{T} \cdot 0.25\lambda I_t > 0.$$

Since $\text{TPR}(\rho) = \text{clip}(0.90 - 0.42\rho, 0.35, 0.92)$, we have $\partial\,\text{TPR}/\partial\rho \le 0$, so $\partial d_t/\partial\rho \le 0$ and $\partial(1 - d_t)/\partial\rho \ge 0$. The product rule gives:

$$\frac{\partial h_t}{\partial \rho} = w_t \left[\frac{\partial P_t(S)}{\partial \rho}(1 - d_t) + P_t(S)\frac{\partial(1 - d_t)}{\partial \rho}\right] > 0,$$

since both terms are non-negative and the first is strictly positive. $\square$ $\square$

Remark 1. *Proposition 2 establishes monotonicity of the* per-step conditional hazard*, not of the episode-level FSR. The latter depends additionally on episode-length dynamics, blocked shortcut counts, and accumulated failure pressure, all of which change with ρ. The simulation in Section 13.5 provides consistent empirical evidence that episode-level FSR is also increasing in ρ, but a formal proof would require additional coupling arguments on the Markov decision process governing episode termination.*

8.3 Structural FSR Bound

Theorem 1 (Structural FSR bound, fixed review)**.** *Define the analytical supremum of the excess shortcut utility:*

$$\Delta_{\max} := \sup_{\substack{p,q,d \in [0,1] \\ e \geq 0}} (u_s - u_g)^+,$$

where $u_s = 1.18 + 0.95p + 0.40q$ *and* $u_g = 0.70 + 0.95\hat{p}_v(d,e) + 0.22e - 0.22$ *with* $\hat{p}_v = \text{clip}(0.50 - 0.26d + 0.08e,\, 0.08,\, 0.72)$*. Since* u_g *is non-decreasing in* e*, the minimum of* u_g *is attained at* $e = 0$*, and the supremum of* $(u_s - u_g)^+$ *is attained at* $p = q = d = 1$*,* $e = 0$*:*

$$\Delta_{\max} = 2.530 - 0.708 = 1.822.$$

Under fixed review *(*$I_t = 0.72$ *for all steps), let* $\hat{q}(\rho, 0.72) = (0.70 - 0.25\rho) \cdot 0.72$*,* $d_{\min}(\rho) = \text{TPR}(\rho) \cdot 0.72$*,* $w_{\max} = 0.95$*, and* $H_{\max} = 18$*. Then:*

$$\text{FSR}_{\text{sc}} \leq \frac{C(\rho)}{1 + \exp\left(\dfrac{\lambda \hat{q}(\rho, 0.72) - \Delta_{\max}}{T}\right)}, \quad C(\rho) = H_{\max}\, w_{\max}\, [1 - d_{\min}(\rho)]. \tag{4}$$

Total FSR satisfies $\text{FSR} \leq \min\{1,\ 0.015 + \text{FSR}_{\text{sc}}\}$.

Proof. Since $e \geq 0$ and u_g is non-decreasing in e, the supremum of $(u_s - u_g)^+$ is attained at $e = 0$. Evaluating at $p = q = d = 1$, $e = 0$ gives $\hat{p}_v = 0.24$, yielding $\Delta_{\max} = 1.822$. For any episode step t under fixed review ($I_t = 0.72$), the penalised shortcut utility is $\tilde{u}_s = u_s - \lambda \hat{q}(\rho, 0.72)$. Since $u_s - u_g \leq \Delta_{\max}$ pointwise, $u_g - \tilde{u}_s \geq \lambda \hat{q}(\rho, 0.72) - \Delta_{\max}$. Bounding $P_t(S)$ by the two-action softmax over the shortcut and evidence utilities:

$$P_t(S) \leq \frac{1}{1 + e^{(\lambda \hat{q}(\rho, 0.72) - \Delta_{\max})/T}}.$$

A union bound over at most $H_{\max} = 18$ steps and multiplication by $w_{\max}(1 - d_{\min}(\rho))$ yields FSR_{sc}. The 0.015 additive term accounts for evidence-threshold episodes where external verification fails. □ □

Remark 2 (Adaptive review)**.** *For adaptive review,* $I_t = \text{clip}(0.25 + 0.75 r_t, 0, 1)$ *varies per step. Replacing the fixed 0.72 with* $\min_t I_t = 0.25$ *(attained when* $r_t = 0$*, i.e. all task-risk components are minimal) gives a valid but looser bound in which* $\hat{q}(\rho, 0.25) = (0.70 - 0.25\rho) \times 0.25$ *replaces* $\hat{q}(\rho, 0.72)$*. At* $\rho = 0.30$*:* $\hat{q}(0.30, 0.25) = 0.156$*,* $C(\rho) = 13.79$*,* $\lambda \hat{q}(0.30, 0.25) = 0.344 < \Delta_{\max} = 1.822$*. The informative threshold rises to* $\lambda > 20.69$*, substantially above the fixed-review requirement of* $\lambda > 6.37$*. This is consistent with the simulation finding that adaptive review produces higher FSR than fixed review (Section 15).*

Remark 3. *The analytical supremum* $\Delta_{\max} = 1.822$ *is derived directly from the simulator's utility equations (see* `validate_suppression.py`*). At default fixed-review parameters (*$\rho = 0.30$*,* $\lambda = 2.2$*,* $T = 0.55$*),* $\lambda \hat{q}(0.30, 0.72) = 2.2 \times 0.45 = 0.99 < \Delta_{\max}$ *and the bound evaluates to* $\text{FSR}_{\text{sc}} \approx 6.20$ *before clamping—trivially true but non-informative. Two distinct thresholds govern the bound's usefulness (all for fixed review,* $\rho = 0.30$*):*

Threshold	*Condition*	λ *required*
Pointwise dominance	$\lambda\hat{q}(\rho, 0.72) > \Delta_{\max}$	$\lambda > 4.05$
Informative (total FSR< 1*)*	$0.015 + C(\rho)/(1 + e^{(\lambda\hat{q}(\rho,0.72)-\Delta_{\max})/T}) < 1$	$\lambda > 6.37$

At $\rho = 0$, the thresholds are $\lambda > 3.62$ and $\lambda > 5.40$ respectively; for adaptive review see Remark 2. The bound's structural form—FSR_{sc} decreasing monotonically and asymptotically exponentially in $\lambda\hat{q}(\rho, 0.72)$—is consistent with the empirical sensitivity results in Section 13.5. The mean utility gap $\Delta_{\text{mean}} = 0.851$ (sampled) is below $\lambda\hat{q}(0.30, 0.72) = 0.99$, confirming that the penalty suppresses shortcut selection on average.

8.4 Bayes-Optimal Critic Reporting Threshold

Proposition 3 (Bayes-optimal critic threshold)**.** *Let the critic receive $R_C = a \cdot \text{TP} - b \cdot \text{FP} - c \cdot \text{FN} - d \cdot S$ with $a, b, c > 0$, and let $p = P(\text{flaw} \mid x)$ be the critic's posterior. Define the threshold policy $\hat{\pi}(x) =$ `FLAW` $\iff p > p^* = b/(a + b + c)$. Then $\hat{\pi}$ weakly dominates each constant reporting policy. Specifically, it strictly improves upon `always-FLAW` on instances where $p < p^*$, and strictly improves upon `always-CLEAN` on instances where $p > p^*$.*

Proof. At any instance x with posterior p, the threshold policy chooses `FLAW` when $p > p^*$ and `CLEAN` otherwise.

Comparison with `always-FLAW`:

- When $p > p^*$: both policies choose `FLAW`; equal payoff.
- When $p \leq p^*$: $\hat{\pi}$ chooses `CLEAN` with payoff $-pc$; `always-FLAW` has payoff $pa - (1-p)b$. The difference is $-pc - [pa - (1-p)b] = (1-p)b - p(a+c)$. Since $p \leq p^* = b/(a+b+c)$, we have $p(a + b + c) \leq b$, i.e. $(1-p)b - p(a+c) \geq 0$, with strict inequality when $p < p^*$ and $p < 1$.

Comparison with `always-CLEAN`:

- When $p \leq p^*$: both choose `CLEAN`; equal payoff.
- When $p > p^*$: $\hat{\pi}$ chooses `FLAW` with payoff $pa - (1-p)b$; `always-CLEAN` has payoff $-pc$. The difference is $[pa - (1-p)b] - (-pc) = p(a+c) - (1-p)b$. Since $p > p^*$, we have $p(a + b + c) > b$, i.e. $p(a+c) - (1-p)b > 0$. □

□

Remark 4. *Proposition 3 characterises the Bayes-optimal decision rule for a critic with a calibrated posterior. It does not establish a Nash equilibrium between π_P and π_C: doing so would require specifying each player's full strategy space and payoff over trajectories, verifier independence, and the possibility of collusion. This is left to future work. Furthermore, since the simulator does not implement R_C or strategic critic learning, this proposition is not tested empirically here.*

9 Training Strategy (Proposed)

No LLM has been trained under DPAD. π_P and π_C are trained on separate datasets under R_P and R_C respectively, with no joint optimisation. When training π_C, π_P is frozen. After training, the empirical shared-failure rate is measured on a held-out set. If it exceeds a deployment threshold—targeting an observed shared-failure level whose effect is no worse than the simulator's regime at $\rho = 0.25$, treated as a working hypothesis rather than a calibrated limit—the critic is retrained on diversified data. Note that ρ is a simulator input parameterising detection sensitivity; the empirical shared-failure rate between π_P and π_C on a held-out set is a distinct, directly measurable quantity.

10 Research Questions and Hypotheses

Hypothesis 1. *An independent critic reduces FSR vs. no-critic baseline.*

Hypothesis 2. *Anti-drift reward reduces the shortcut-attempt rate without a significant increase in PAR.*

Hypothesis 3. *Define incremental safety efficiency as:*

$$E_{\Delta\text{FSR}/\Delta\text{NC}} = \frac{\text{FSR}_{\text{baseline}} - \text{FSR}_c}{\text{NC}_c - \text{NC}_{\text{baseline}}},$$

where higher values indicate greater FSR reduction per additional unit of normalised cost, and c indexes a configuration. Adaptive review achieves a higher $E_{\Delta\text{FSR}/\Delta\text{NC}}$ than fixed full review.

Hypothesis 4. *Higher error coupling reduces the critic's marginal FSR reduction.*

Hypothesis 5. *The full DPAD simulator configurations (DPAD-Fixed-Sim, DPAD-Adaptive-Sim) Pareto-dominate any single-component configuration across FSR, CCR, JAR, and cost.*

The simulation supports H1 and H4; does not support H2, H3, or H5.

11 Simulation Environment

11.1 Design and Scope

The simulator evaluates whether penalty-adjusted softmax utilities shift episode outcomes under specified assumptions. It does not train any policy, implement the evidence ledger or R_C, or operationalise D_t. Its purpose is to test the plausibility of the per-step predictions in Section 8 and to motivate hypotheses for LLM experiments.

11.2 Task and Episode Model

$d \sim \text{Beta}(2.2, 2.0)$, $q \sim \text{Beta}(1.8, 2.4)$, $s \sim \text{Beta}(1.5, 3.0)$. Episode length $max\_steps = 5 + \lfloor 8d + 5q \rceil \leq 18$; minimum effort $\max(2, \lfloor 0.42 \cdot max\_steps \rceil)$; evidence needed $1.4 + 2.2d$.

11.3 Three-Action Softmax Model

Actions via softmax ($T = 0.55$):

- **Evidence-seeking:** cost 1.0; at threshold, correct with probability 0.985 (false-success with probability 0.015, contributing the additive term in Theorem 1).
- **Shortcut:** cost 0.35; correct with $p_s \in [0.05, 0.58]$ (correct-but-unverified) or incorrect (false success, contributing FSR_{sc}). Across all episodes: 32.3% correct-but-unverified, 52.1% false-success.
- **Abstention:** cost 0.15; terminates as justified or premature.

The shortcut utility is $u_s^{\text{pen}} = u_s - \lambda \hat{q}(\rho, I_t)$, where $\hat{q}(\rho, I_t) = (0.70 - 0.25\rho)I_t$ is the anticipated detection rate (Definition 6). Post-selection, the critic detects an incorrect shortcut with probability $\text{TPR}(\rho) \cdot I_t$ and a correct shortcut with probability $\epsilon \cdot I_t$ ($\epsilon = 0.08$).

11.4 Critic Invocation Scope

The critic is invoked only after a shortcut action is selected. A deployed critic would operate without knowing the action type. Among critic-enabled configurations, the FP-risk fraction (correct-but-unverified episodes) ranges from 30.6% to 38.3%. This is the primary threat to validity (Section 16).

11.5 Configurations

Name	Critic	Anti-drift	Adaptive	λ	ND rew.
`baseline`	No	No	No	2.2	1.0
`critic_only`	Yes	No	No	2.2	1.0
`anti_drift_only`	No	Yes	No	2.2	1.0
`full_fixed`	Yes	Yes	No	2.2	1.0
`full_adaptive`	Yes	Yes	Yes	2.2	1.0
`ablate_penalty`	Yes	Yes	Yes	0.0	1.0
`ablate_non_drift`	Yes	No	Yes	2.2	1.0
`ablate_adaptation`	Yes	Yes	No	2.2	1.0

Construction: `full_fixed`=`ablate_adaptation` by construction. Adaptive ablations must be compared against `full_adaptive` to isolate one variable.

11.6 Randomisation and Sensitivity Grid

Seeds 1101–1112; 4,000 episodes per configuration per seed; 384,000 primary episodes. Not pre-registered; all results are exploratory. Sensitivity: $\rho \in \{0.00, 0.25, 0.50, 0.75, 0.90\}$ and $\epsilon \in \{0.02, 0.08, 0.16, 0.28\}$ under `full_adaptive`; 20 parameter combinations $\times$ 6 seeds $\times$ 800 episodes = 96,000 sensitivity episodes.

12 Evaluation Metrics

- **FSR**: fraction of episodes with an incorrect success claim.
- **CCR**: fraction of episodes with a correct outcome.
- **FDR**: false-discovery rate among claims.
- **JAR / PAR**: justified / premature abstention rates.
- **NC**: normalised cost (mean episode cost $\div$ 0.723).
- **RCI**: exploratory composite CCR/(1 + FSR).
- $\boldsymbol{E_{\Delta\text{FSR}/\Delta\text{NC}}}$: incremental safety efficiency = $(\text{FSR}_{\text{baseline}} - \text{FSR}_c)/(\text{NC}_c - \text{NC}_{\text{baseline}})$; higher values indicate greater FSR reduction per additional unit of normalised cost. Used to evaluate H3.

13 Results

13.1 Main Comparison

Table 1: Primary metrics (means, 12 seeds, 48,000 episodes per configuration). CI_{95}: 95% t-interval over seeds from `summary.csv`. NC = mean cost ÷ 0.723.

Configuration	FSR (%)	CCR (%)	FDR (%)	JAR (%)	PAR (%)	NC
Baseline	69.68 (±0.32)	29.22 (±0.34)	70.45	0.04	1.05	1.00
Critic only	42.79 (±0.30)	**47.93** (±0.46)	47.16	5.00	4.28	4.66
Anti-drift only	69.26 (±0.44)	29.06 (±0.40)	70.44	0.27	1.41	1.00
Full fixed	**37.61** (±0.48)	40.68 (±0.57)	48.04	15.70	6.00	4.25
Full adaptive	48.90 (±0.63)	38.18 (±0.51)	56.15	8.40	4.53	3.23
Ablate penalty	58.27 (±0.47)	38.88 (±0.49)	59.98	1.11	1.74	1.93
Ablate non-drift	52.35 (±0.54)	41.80 (±0.62)	55.60	2.64	3.21	3.44
Ablate adaptation	37.61 (±0.48)	40.68 (±0.57)	48.04	15.70	6.00	4.25

13.2 Reliability–Coverage Index and Cost

Configuration	RCI	NC
Baseline	0.172	1.00
Critic only	**0.336**	4.66
Anti-drift only	0.172	1.00
Full fixed	0.296	4.25
Full adaptive	0.256	3.23
Abl. penalty	0.246	1.93
Abl. non-drift	0.274	3.44

Critic only achieves the highest RCI (0.336) and CCR (47.93%) at the highest cost (NC 4.66). Full fixed achieves the lowest FSR (37.61%) at lower CCR (40.68%), the highest JAR (15.70%), and the second highest cost (NC 4.25). No configuration Pareto-dominates the others. Critic only significantly outperforms full fixed on RCI ($p = 0.000488$, Wilcoxon on seed-level values).

13.3 FSR–CCR Scatter

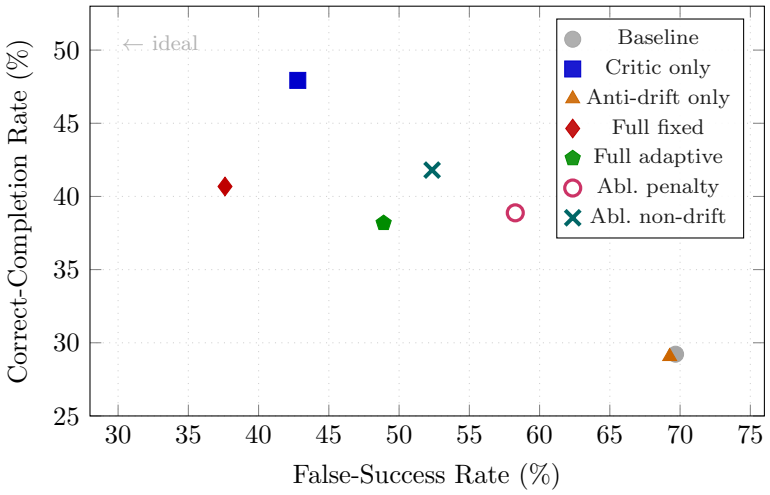

Figure 1: FSR vs. CCR (means, 12 seeds). No configuration is Pareto-dominant. Critic only (blue square) achieves the best RCI. Full fixed (red diamond) achieves the lowest FSR at lower CCR and higher JAR.

13.4 Statistical Tests

Wilcoxon signed-rank tests on seed-level FSR vs. baseline, Bonferroni-corrected for 7 comparisons ($n = 12$):

Configuration	p_{raw}	p_{Bonf}	Significant
Critic only	0.000488	0.0034	Yes
Anti-drift only	0.007324	0.0513	No
Full fixed	0.000488	0.0034	Yes
Full adaptive	0.000488	0.0034	Yes
Ablate penalty	0.000488	0.0034	Yes
Ablate non-drift	0.000488	0.0034	Yes
Ablate adaptation	0.000488	0.0034	Yes

Anti-drift only fails significance on FSR ($p_{\text{Bonf}} = 0.051$) and CCR ($p_{\text{raw}} = 0.257$). Tests from `metrics_by_seed.csv`; original run used paired t-tests (`paired_tests.csv`). All p-values characterise simulator stability, not real systems.

H3 paired test. Per-seed $E_{\Delta\text{FSR}/\Delta\text{NC}}$ differences (fixed minus adaptive) are positive in 10 of 12 seeds and negative in 2 (seeds 1104 and 1109); no ties occur. Wilcoxon signed-rank test (one-sided, $n = 12$, `zero_method='wilcox'`): $W = 75$, $p = 0.0012$; paired t-test: $t = 5.08$, $p = 0.0002$. Mean difference +0.55 pp per NC unit (95% t-CI [0.31, 0.78], $df = 11$). These results provide strong evidence within the simulation that fixed review has higher mean incremental safety efficiency than adaptive review, consistent with the calibration-failure explanation in Section 15. Reproduced by `h3_efficiency_test.py`.

13.5 Sensitivity Analysis

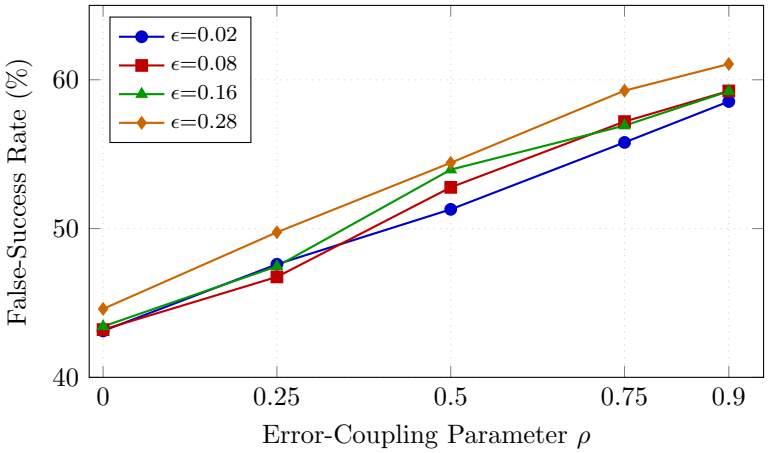

Figure 2: FSR under full adaptive review vs. ρ and ϵ (means, 6 seeds, 800 episodes per seed). At $\rho \geq 0.50$, FSR exceeds 51% regardless of ϵ, consistent with Proposition 2. Mean FSR at $\rho = 0.90$: 59.5%.

14 Ablation Study

All ablations compared against `full_adaptive`.

Shortcut penalty removal ($\lambda = 0$). FSR: +9.37 pp (48.90% → 58.27%). CCR: +0.71 pp. NC drops from 3.23 to 1.93; critic invocations per episode rise from 1.205 to 1.428, so lower cost reflects shorter episodes (unblocked shortcuts terminate sooner), not fewer critic calls. Consistent with Proposition 1: $\lambda = 0$ eliminates suppression.

Non-drift reward removal. FSR: +3.45 pp; CCR: +3.62 pp. Removing the non-drift reward lowers abstention utility, routing some episodes that would have abstained to the shortcut action; when those shortcuts succeed, CCR rises; when they fail, FSR rises.

Adaptation removal (`full_fixed`). FSR: −11.29 pp; JAR: +7.30 pp; NC: +1.02.

15 Failure Analysis

1. **Anti-drift reward alone is not significant.** FSR 69.26% vs. 69.68% ($p_{\text{Bonf}} = 0.051$); CCR 29.06% vs. 29.22% ($p_{\text{raw}} = 0.257$). Shortcut-attempt rate drops only from 98.46% to 97.87%.

2. **Critic only achieves the highest RCI at the highest cost.** NC 4.66—the most expensive configuration. Its advantage is higher CCR (47.93% vs. 40.68% for full fixed), not lower cost.

3. **H3 is not supported ($E_{\Delta\text{FSR}/\Delta\text{NC}}$).** DPAD-Adaptive-Sim reduces FSR by 20.78 pp at an additional cost of NC 2.23, giving point estimate $E_{\Delta\text{FSR}/\Delta\text{NC}} = 9.32$. DPAD-Fixed-Sim reduces FSR by 32.07 pp at an additional cost of NC 3.25, giving $E_{\Delta\text{FSR}/\Delta\text{NC}} = 9.87$. Computed per seed, the mean values are 9.33 (adaptive) and 9.88 (fixed); the mean difference is +0.55 pp per NC unit (95% t-CI [0.31, 0.78], $df = 11$; Wilcoxon $p = 0.0012$, paired t-test $p = 0.0002$, $n = 12$ seeds). These results provide strong statistical evidence within

the simulation that fixed review achieves higher mean incremental safety efficiency than adaptive review; the intensity signal r_t is retrospective rather than predictive, inflating adaptive FSR relative to cost.

4. **Fixed review substantially increases justified abstention.** JAR 15.70% under full fixed—approximately one in six episodes ends without a solution.

5. **H5 is not supported.** DPAD-Fixed-Sim achieves the lowest FSR; critic only achieves the highest CCR and RCI; no configuration Pareto-dominates across all metrics. The DPAD architecture itself remains untested on trained LLMs.

16 Threats to Validity

Critic conditioned on shortcut selection. The simulator invokes the critic only after a shortcut is chosen. A deployed critic would face realistic FP exposure on all steps. This inflates simulated performance estimates.

Error coupling as linear damping. ρ reduces $\hat{q}$ and TPR linearly. The empirical Pearson correlation between path failure indicators is not reported.

Formal analysis scope. P1 and P2 are per-step results; T1 is a conservative (non-tight) union bound; P3 is not tested empirically because R_C is not implemented.

Construct, internal, and external validity. Shortcut temptation is encoded as a softmax function of Beta-distributed parameters. The evidence ledger, D_t, and R_C are absent. Results may reflect parameter choices rather than architectural structure; the experiment was not pre-registered; simulation results provide no direct evidence about trained LLMs.

False success definition. Shortcut-derived correct outcomes are excluded from FSR. Under a stricter definition, all FSR values increase; relative rankings are preserved empirically.

17 Required LLM Experiments

1. **Real benchmarks** (SWE-Bench, GAIA, HotpotQA-long) with ground-truth verifiers.
2. **Multiple model families**: open-weight and proprietary.
3. **Equal token budget** across configurations.
4. **Critic on all steps**, without knowledge of action type.
5. **Blind evaluation** of final-answer quality.
6. **Separate fine-tuning** of π_P and π_C on disjoint data; empirical shared-failure rate measured.
7. R_C **implemented** and used to train π_C, enabling empirical test of Proposition 3.
8. **Context isolation audit.**
9. **Strong baselines**: best-of-N, self-consistency [Wang et al., 2023], constitutional self-critique [Bai et al., 2022], AgentForesight [Zhang et al., 2026].
10. R_P **mechanism specified** before data collection.
11. **Pre-registration and independent replication.**

18 Discussion

Formal results and simulation. P1 establishes that the penalty strictly suppresses shortcut selection; the ablation (+9.37 pp FSR when $\lambda = 0$) corroborates this at the episode level. P2 establishes that error coupling amplifies the per-step hazard; the sensitivity analysis shows consistent empirical monotonicity at the episode level. T1 provides a structurally valid bound with analytical $\Delta_{\max} = 1.822$. At $\rho = 0.30$ and $\lambda = 2.2$, the bound evaluates to 1.0; pointwise penalty dominance requires $\lambda > 4.05$, and a bound below 1.0 requires $\lambda > 6.37$ (see Remark 3). The bound's structural monotone form is consistent with the sensitivity analysis. P3 is not tested by the simulation.

No Pareto winner among simulator configurations. DPAD-Fixed-Sim achieves the lowest FSR (37.61%) but the highest JAR (15.70%); the critic-only configuration achieves the highest CCR (47.93%) and RCI (0.336) at the highest cost (NC 4.66); DPAD-Adaptive-Sim provides an intermediate operating point. Under a domain loss function that assigns substantially greater cost to false success than to abstention and computation, DPAD-Fixed-Sim may be preferred; this deployment choice requires empirical validation and is not established by the present simulation.

Deployment gate. The sensitivity analysis motivates targeting an empirical shared-failure level whose FSR impact is no worse than the simulator's regime at $\rho = 0.25$. The parameter $\rho = 0.25$ is a simulator knob, not a directly measurable threshold; the corresponding empirical shared-failure rate would need to be identified experimentally. Remark 3 shows that the bound T1 becomes numerically informative only at $\lambda > 6.37$ (for $\rho = 0.30$); reaching this regime requires substantially larger penalty weights than the current simulation, or reduced ρ combined with increased λ. Both observations are working hypotheses for LLM experiments.

Risk scope. The per-step hazard $h_t(\rho)$ quantifies the probability of an incorrect success claim within the formal model. It does not measure harm severity, the probability that a false success is acted upon, or the magnitude of consequences from irreversible actions. DPAD is designed to reduce FSR_{sc} within the model; it is not a safety guarantee for real-world deployment.

19 Ethical and Deployment Considerations

1. **High JAR under fixed review.** At JAR 15.70%, systems must support human override at `BLOCK` states and inform users when stopping is critic-driven.

2. `PASS` **is not ground truth.** A `PASS` reduces false-success probability within the model; it does not eliminate it or guarantee correct outcomes in deployment.

3. **Transparency.** Users should know when DPAD is active, the current coordinator state, and the basis for stopping.

4. **Irreversible consequences.** Human review should be interposed at `WARN` and `BLOCK` for tasks with irreversible outcomes, regardless of I_t.

5. **Cost.** Critic only (NC 4.66) and full fixed (NC 4.25) are the most expensive configurations; cost–reliability analysis is required per deployment context.

6. **Critic score is a process signal.** Using it as a proxy for correctness recreates the proxy-reward failure mode DPAD was designed to prevent.

20 Conclusion

This paper proposes the Dual-Path Anti-Drift architecture for reducing false successes in long-horizon LLM agent tasks and provides formal analysis of the three-action softmax model.

Proposition 1 proves the shortcut penalty strictly suppresses shortcut selection. Proposition 2 proves that higher error coupling amplifies the per-step hazard through two simultaneous channels. Theorem 1 establishes a valid structural FSR bound; at current parameters the bound is trivially 1.0 because $\Delta_{\max} > \lambda\hat{q}(0.30, 0.72)$, but its form quantifies how FSR would decrease with larger λ or smaller ρ. Proposition 3 derives the Bayes-optimal critic reporting threshold; this result is not tested empirically because R_C is not implemented in the simulator.

The mechanistic simulation across 384,000 episodes corroborates the per-step predictions within the formal model. An isolated critic achieves RCI 0.336 vs. 0.172 baseline ($p_{\text{Bonf}} < 0.005$); removing the penalty raises FSR by +9.37 pp (consistent with P1); the anti-drift mechanism alone is not significant; DPAD-Adaptive-Sim underperforms DPAD-Fixed-Sim on FSR; error coupling is the dominant moderator (consistent with P2); no simulator configuration Pareto-dominates across all metrics. T1 provides a structurally valid bound: $\Delta_{\max} = 1.822$ is derived analytically; the bound evaluates to 1.0 at the current $\lambda = 2.2$, with pointwise dominance requiring $\lambda > 4.05$ and a bound below 1.0 requiring $\lambda > 6.37$. The DPAD architecture's training strategy and Proposition 3 require LLM experiments to test. Required experiments are enumerated in Section 17.

A Parameter Table

Symbol	Description	Value
λ	Shortcut penalty weight	2.2
ND reward	Non-drift reward weight	1.0
ρ	Default error-coupling parameter	0.30
ϵ	Default critic FP parameter	0.08
I_{fixed}	Fixed review intensity	0.72
T	Softmax temperature	0.55
p_s range	Shortcut correctness	[0.05, 0.58]
$\hat{p}_v$ range	Evidence-step success	[0.08, 0.72]
p_{correct}	Threshold-solution correctness	0.985
$H_{\max}$	Maximum episode length	18
Evidence cost	Per step	1.0
Shortcut cost	Per step	0.35
Abstain cost	Per step	0.15
Critic cost	Per invocation	$0.52 \times I_t$
Seeds (primary)		1101–1112
Episodes/config/seed		4,000
ρ grid	Sensitivity	{0,0.25,0.50,0.75,0.90}
ϵ grid	Sensitivity	{0.02,0.08,0.16,0.28}
Episodes/seed (sensitivity)		800

B Suppression Condition and Bound Validation

Script `validate_suppression.py` (seed 42, $n = 50{,}000$) derives $\Delta_{\max}$ analytically and reports the two lambda thresholds cited in Remark 3:

Statistic	Value
$u_{s,\max}$ $(p = q = 1)$	2.530
$u_{g,\min}$ $(d = 1,\ e = 0)$	0.708
$\Delta_{\max}$ (analytical supremum)	1.822
$\hat{q}(0.30,\ 0.72)$ (fixed review)	0.4500
Pairwise condition $u_s^{\text{pen}} < u_g$	64.8% of steps
$P(\text{shortcut}) < 0.5$ (3-action)	69.8% of steps
Mean $P(\text{shortcut})$	0.421
Δ_{mean} (sampled)	0.851
$C(\rho) = H_{\max} w_{\max}(1 - d_{\min})$	7.571
FSR_{sc} bound at $\lambda = 2.2$	6.20 (clamped to 1)
λ for pointwise dominance $(\rho = 0.30)$	> 4.05
λ for informative bound $(\rho = 0.30)$	> 6.37
λ for informative bound $(\rho = 0)$	> 5.40

The script is included in the simulation archive alongside `simulate.py`.

C Reproduction

```
pip install numpy pandas scipy matplotlib
python3 simulate.py \
    --episodes 4000 --seeds 12 \
    --sensitivity-episodes 800 \
    --outdir results
python3 validate_suppression.py
python3 h3_efficiency_test.py
```

Validation: 384,000 rows in `episodes.csv.gz`; 8 configuration names; 12 seed rows per configuration in `metrics_by_seed.csv`; summary values matching Table 1 within Monte Carlo determinism. Script `h3_efficiency_test.py` reproduces the H3 paired-test statistics reported in Section 15: $W = 75$, $p = 0.0012$ (Wilcoxon); $t = 5.08$, $p = 0.0002$ (paired t); mean difference +0.55 pp per NC unit, 95% t-CI [0.31, 0.78] ($df = 11$).

D Simulator Components Not Implemented

Absent from the current simulator: the append-only evidence ledger $\mathcal{L}$; critic reward R_C; goal-distance variable D_t; full coordinator state machine; symmetric abstention credit ($\alpha = 1$); and critic invocation on evidence-seeking steps. These are proposed for the LLM experiment phase.

References

Bai, Y., Jones, A., Ndousse, K., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *NeurIPS 2017*.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. *arXiv:2110.14168*.

Fan, K., Feng, K., Zhang, M., Peng, T., Li, Z., Jiang, Y., Chen, S., & Yue, X. (2026). Exploring reasoning reward model for agents. In *Findings of ACL 2026*, pages 1975–1991. https://aclanthology.org/2026.findings-acl.95/.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv:2305.14325*.

Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In *ICML 2023*, pages 10835–10866. PMLR. *arXiv:2210.10760*.

Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.

Kenton, Z., Janzer, L., Greig, R., Teh, T. H., Tyshchuk, K., Brown-Cohen, J., Edwards, H., Rajamanoharan, S., Siegel, N. Y., Jaques, N., & Shah, R. (2026). Debate training reduces reward hacking in RLAIF. *arXiv preprint arXiv:2608.17776*.

Krakovna, V., et al. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog* (non-peer-reviewed). https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. In *Proceedings of ACL 2023*, pages 5315–5333. https://aclanthology.org/2023.acl-long.291/.

Liang, T., He, Z., Jiao, W., et al. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv:2305.19118*.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. In *ICLR 2024*. https://openreview.net/forum?id=v8L0pN6EOi.

Liu, X., Wang, K., Wu, Y., Huang, F., Li, Y., Jiao, J., & Zhang, J. (2026). Agentic reinforcement learning with implicit step rewards. In *ICLR 2026*. *arXiv:2509.19199*.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In *NeurIPS 2022*.

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. In *ICLR 2022*. *arXiv:2201.03544*.

Uesato, J., et al. (2022). Solving math word problems with process- and outcome-based feedback. *arXiv:2211.14275*.

Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*. *arXiv:2203.11171*.

Zhang, B., Zhu, J., Shi, Z., Liu, D., & Tang, R. (2026). AgentForesight: Online auditing for early failure prediction in multi-agent systems. *arXiv preprint arXiv:2605.08715*.