Dual-Path Anti-Drift (DPAD) Architecture

Scientific Review, Scoring, and Publication Roadmap

A structured critique, comparative score table, and improvement plan for submission to a peer-reviewed venue

---

Author:	Yuval Fradkin
Date:	September 20, 2026
Original paper:	Fradkin, Y. (2026). *Dual-Path Anti-Drift Reward Architecture for Reliable Long-Horizon LLM Agents: A Mechanistic Simulation Study with Formal Analysis.*
Review scope:	Scientific quality, scoring, simulation fixes, venue guidance, and references

Abstract

Long-horizon language model agents face a structural incentive to report false successes rather than continue gathering evidence. The Dual-Path Anti-Drift (DPAD) architecture addresses this by separating the primary solver from an independent critic under a distinct reward function, supported by a deterministic coordinator and an append-only evidence ledger. This document provides a structured scientific review of the DPAD paper (Fradkin, 2026), covering eight evaluation dimensions, four formal remarks that correct methodological errors in earlier versions of this review, a prioritised list of required simulation fixes, and a bibliography of 23 references. The central finding is that DPAD's current simulation contains one critical design limitation (the critic is conditioned on action type unavailable at deployment) and one near-tautological coupling model, both of which can be corrected without LLM experiments. After those corrections, together with budget-matched baselines and a pre-registered confirmatory run, the paper would become a strong formal simulation study with a projected peer-review readiness score of 7.5–8.0 out of 10.

Contents

Comparative Score Table

Scoring conventions

All dimensions are scored on a 1–10 scale. Three states are distinguished:

> Scores under **S1 (Current)** reflect the paper as submitted. Scores under **S3 (After implementation)** are projections conditional on the simulation fixes described in Sections 3–4. No score reflects any new data, simulation run, or proof not already present in the paper.

S1 — Current.
The paper as submitted (September 2026), before any revision.

S2 — After this review document.
The review document adds no new data, simulation run, or proof. Only scientific transparency changes: the design–deployment gap is now explicit and public.

S3 — After implementation and verification.
Projected scores conditional on C1, C2, I1, I2, I3 being fully executed and validated. These are estimates, not guarantees.

Scores for the DPAD paper

Dimension	S1	S2	S3	Basis for S3 change
Scientific quality (overall)	6.5	6.5	7.8	Critic on all steps; latent-variable coupling; baselines
Architectural novelty	6.8	6.8	7.5	Isolation contract sharpened by budget-matched ablations
Practical novelty	7.4	7.4	8.2	Deployment-representative simulation; cost–reliability table
Mathematical formality	6.5	6.5	7.2	P2 episode conjecture explicit; T1 non-informativeness in body
Simulation-specific empirical validity	3.8	3.8	7.0	Full-step critic; causal latent-variable ρ; pre-registered run
Scientific transparency	9.2	9.5	9.5	Roadmap makes design–deployment gap explicit
Peer-review readiness	5.5	5.5	7.5–8.0	All Critical and Important fixes executed
Preprint readiness	8.0	8.5	9.0	Structured disclosure alongside the preprint
Unweighted mean	**6.5**	**6.6**	**8.0**	

Score notes

Mathematical formality (S1 = 6.5, not 8.0).
Within the formal model, all four proofs are internally correct and proof correctness alone merits 8/10. As a score of overall *mathematical contribution*, 6.5 is appropriate: P1 and P2 follow from the softmax Jacobian and chain rule applied to the model's own utility definitions; P3 is a known Bayes-optimal threshold result for a linear utility function [4]; and T1 is a conservative union bound that evaluates to 1.0 (non-informative) at current parameters.

Simulation-specific empirical validity (S1 = 3.8).
The label "simulation-specific" is deliberate. The score of 7.0 in S3 applies only within the corrected simulation model. It does not constitute empirical validity for DPAD as an architecture for trained LLMs, which remains untested without the LLM experiments listed in Section 17 of the paper.

Peer-review readiness and main-track benchmarks.
A single focused LLM experiment on one benchmark with one open-weight model would materially strengthen suitability for main-track venues such as NeurIPS, ICML, or ICLR. Acceptance also depends on novelty, baseline strength, and results; no experiment alone guarantees admission.

Four Methodological Corrections

Remark 1 (Direction of bias in C1)**.** An earlier formulation stated that all FSR, CCR, RCI, and cost figures are "optimistic estimates relative to a real deployment." This is not established without an empirical check.

Invoking the critic on every step exposes it to false-positive risk on evidence-seeking and abstention steps. Depending on the critic's FP rate on non-shortcut steps, FSR could *increase* (if legitimate actions are blocked) or *decrease* further (if the critic catches flaws missed by post-selection review). Cost increases unambiguously.

Corrected formulation: The current estimates are not deployment-representative because the critic is conditioned on action type unavailable at inference time. The direction of bias must be determined empirically.

Remark 2 (Best-of-N selection mechanism)**.** Ground-truth selection across N independent episodes is an oracle procedure, not a fair comparison baseline [21]. It assumes knowledge of which episodes succeed before selecting among them, which is precisely what the system is designed to determine.

Required split.

(a) **Oracle upper bound.** N episodes; ground-truth selection. Reported as a ceiling, not as a fair comparison.

(b) **Primary budget-matched baseline.** N episodes; a fixed verifier with stated accuracy and cost selects the best outcome without access to ground truth [3]. N is set so that total cost equals NC = 4.66 (critic-only).

Remark 3 (Causal latent-variable model for C2)**.** The latent-variable model proposed in earlier versions introduced $Z \sim \mathcal{N}(0,1)$ as a shared factor but placed it in both paths with the same sign:

$$\tilde{u}_{s,t} = u_s - \lambda \hat{q}(\rho, I_t) + \gamma_P Z, \qquad \mathrm{TPR}_t = \mathrm{clip}(0.90 - 0.42\rho + \gamma_C Z,\, 0.35,\, 0.92).$$

This formulation has two errors. First, if Z represents a shared difficulty or blind-spot, a positive Z (harder episode) should increase shortcut temptation *and* decrease detection ability, producing a negative contribution to TPR, not a positive one. The signs are inconsistent. Second, ρ remains in both equations with the same linear-damping form as before, so the tautological coupling problem (Section C2 below) is not eliminated.

Corrected model. Remove ρ from the direct coupling channel. Let Z be a per-episode shared difficulty factor with *opposite signs* in the two paths:

$$\tilde{u}_{s,t} = u_s - \lambda \hat{q}_0(I_t) + \gamma_P Z \tag{1}$$

$$\mathrm{TPR}_t = \mathrm{clip}(\mathrm{TPR}_0(I_t) - \gamma_C Z,\ 0.35,\ 0.92) \tag{2}$$

where $\hat{q}_0(I_t)$ and $\mathrm{TPR}_0(I_t)$ are baseline functions independent of Z, and $\gamma_P, \gamma_C \geq 0$. A positive Z increases shortcut temptation (primary path less reliable) and decreases detection rate (critic less reliable): both failure modes move in the same direction, producing genuine shared-failure correlation. The empirical ϕ-coefficient or mutual information $I(F_P; F_C)$ then measures an emergent property, not a construction artefact. If a sensitivity parameter analogous to ρ is still desired, it should enter as a separate multiplier on γ_P and γ_C, not as a direct component of the utility or TPR formulas.

Remark 4 (Scope of C1 impact on formal results)**.** C1 requires re-running the simulation with the critic on every step. This does not invalidate all four formal results.

- **P1** (monotone shortcut suppression) depends on $\partial P(S)/\partial\lambda$ and the softmax structure. These are independent of when the critic is invoked. P1 remains valid unless the utility model changes.
- **P3** (Bayes-optimal critic threshold) depends on R_C and the critic's posterior. It is independent of invocation scope. P3 remains valid.
- **P2** (per-step hazard amplification) and **T1** (structural FSR bound) both depend on $\mathrm{TPR}(\rho)$. If the full-step critic has a different effective TPR function under the corrected model (equations 1–2), P2 and T1 must be re-derived.

Critical Issues

C1 — Critic invoked only after shortcut selection

Description. The simulator activates the critic only after a shortcut action is chosen (Section 11.4 of the paper). A deployed critic operates on every step without advance knowledge of the action type and faces false-positive risk on evidence-seeking and abstention steps.

Consequence. All metrics in Table 1 are not deployment-representative (Remark 1). P2 and T1 require re-derivation if the full-step TPR function differs (Remark 4).

Required fix.

(1) Re-run all configurations with the critic on every step, without conditioning on action type.

(2) Report FP rate on evidence-seeking steps, FP rate on abstention steps, TPR on shortcut steps, and NC change.

(3) Present a side-by-side table of all Table 1 metrics before and after the change, with 95 % CIs across seeds.

(4) If TPR changes, re-derive P2 and T1 under the new function.

C2 — Error-coupling model is near-tautological and incorrectly signed

Description. The parameter ρ directly reduces both $\hat{q}(\rho, I_t)$ and $\mathrm{TPR}(\rho)$ by linear damping. Because ρ is wired into both, the correlation between failure indicators is determined by the model input, not an emergent property of the simulation. The corrected latent-variable model (Remark 3) also requires opposite signs: a harder episode ($Z > 0$) must increase shortcut temptation *and* decrease detection ability.

Required fix. Implement equations (1)–(2). After implementation:

(a) Report the empirical ϕ-coefficient [16] or mutual information $I(F_P; F_C)$. For binary indicators, ϕ is preferred over Pearson r.

(b) Present a sensitivity grid over (γ_P, γ_C).

(c) Clearly distinguish design parameters from emergent coupling.

Important Improvements

I1 — Implement R_C and test Proposition 3

Proposition 3 derives the Bayes-optimal critic threshold $p^* = b/(a+b+c)$ under $R_C = a \cdot \mathrm{TP} - b \cdot \mathrm{FP} - c \cdot \mathrm{FN} - d \cdot S$. The simulator does not implement R_C and the proposition is not empirically evaluated.

A genuine test requires: (a) a calibrated posterior score $p \in [0, 1]$ on a held-out calibration set, with a reliability (calibration) curve and Brier score or Expected Calibration Error (ECE) reported; (b) a threshold decision at p^* as a function of (a, b, c); (c) comparison of the threshold policy, always-FLAW, and always-CLEAN on held-out episodes with known ground truth; and (d) reporting of expected R_C under each policy with 95 % CIs. A fitted distribution score is not calibrated by construction; calibration must be measured on a separate held-out set.

I2 — Budget-matched baselines

(a) **Oracle upper bound.** Best-of-N with ground-truth selection at NC$=$ 4.66. Reported as ceiling, not fair comparison (Remark 2).

(b) **Primary budget-matched baseline.** Best-of-N with a fixed verifier at equal cost. Verifier accuracy and cost stated.

(c) **Shared-context critic.** A critic sharing the primary path's context window and not isolated in reward. This is the direct ablation of DPAD's structural isolation claim.

(d) **Equal-cost reporting.** FSR, CCR, RCI for all configurations at equal NC on a common axis.

I3 — Pre-registration of confirmatory run

(1) Apply C1 and C2 fixes fully before freezing parameters. The corrected TPR function may change the sensitivity grid and the statistical plan; freeze only after C1/C2 are finalised.

(2) File a pre-registration on OSF (`https://osf.io`) or AsPredicted (`https://aspredicted.org`), specifying seeds, episodes per configuration, primary outcome, and the planned test ($\alpha = 0.05$, Bonferroni-corrected).

(3) Run the confirmatory simulation after the time-stamped registration is locked.

(4) Report exploratory (current) and confirmatory results in separate tables.

Minor Issues

M1.
Shorten the abstract to 200–250 words.

M2.
Merge Section 2 (Terminology Conventions) into Section 4 (Terminology and Problem Formulation); place Related Work immediately after the Introduction.

M3.
Add horizontal and vertical 95 % CI error bars to Figure 1. The values already exist in Table 1.

M4.
Move the unimplemented-components list (Appendix D) to Section 16 (Threats to Validity) as a numbered sub-list. Appendix D may retain a technical description.

M5.
State T1 non-informativeness in the main body of Section 8.3:

> *At current parameters ($\lambda = 2.2$, $\rho = 0.30$, fixed review), $\lambda\hat{q}(0.30, 0.72) = 0.99 < \Delta_{\max} = 1.822$ and the bound provides no quantitative constraint. It becomes informative only at $\lambda > 6.37$.*

Priority Order

Priority	Item	Action	Class
1	C1	Full-step critic; no action-type conditioning	Critical
2	C2	Causal latent-variable model (eqs. 1–2); opposite signs; ρ removed from direct coupling; report ϕ or $I(F_P; F_C)$	Critical
3	I1	Calibrated posterior (reliability curve + ECE/Brier); threshold rule; empirical P3 test	Important
4	I2	Oracle upper bound + verifier-based primary baseline; shared-context critic ablation	Important
5	I3	Pre-register after C1/C2 are finalised	Important
6	M2	Merge §2 into §4; reorder sections	Minor
7	M4	Unimplemented components to §16	Minor
8	M1	Shorten abstract	Minor
9	M3	Error bars on Figure 1	Minor
10	M5	T1 non-informativeness in §8.3 body	Minor

Formal Analysis Notes

P1 (Monotone shortcut suppression).
Correct. $\partial P(S)/\partial\lambda < 0$ follows from the softmax Jacobian. The ablation (+9.37 pp FSR at $\lambda = 0$) corroborates it at the episode level. P1 is unaffected by C1 (Remark 4).

P2 (Per-step hazard amplification).
Correct for the per-step hazard $h_t(\rho)$. Remark 1 of the paper correctly labels the episode-level FSR monotonicity as empirically supported but not formally proved. After C1: if the corrected full-step TPR function differs, P2 must be re-derived. Either a coupling argument on the governing MDP or an explicit conjecture label is needed.

T1 (Structural FSR bound).
Analytically valid. $\Delta_{\max} = 1.822$ is correct. At default parameters: $\lambda\hat{q}(0.30, 0.72) = 2.2 \times 0.45 = 0.99 < \Delta_{\max}$. The bound evaluates to $\mathrm{FSR}_{sc} \approx 6.20$, clamped to 1.0: trivially true and uninformative. Informative only at $\lambda > 6.37$ (fixed review, $\rho = 0.30$). After C1/C2: T1 must be re-derived if TPR changes. This must be stated prominently in the body (M5).

P3 (Bayes-optimal critic threshold).
Correct. $p^* = b/(a+b+c)$ weakly dominates constant strategies. Not empirically tested. Unaffected by C1 (Remark 4). See I1.

Key Empirical Finding

The most practically significant result is the critic-only versus full-fixed trade-off:

Configuration	FSR	CCR	RCI	JAR	NC
Critic only	42.79 %	**47.93 %**	**0.336**	5.00 %	4.66
Full fixed	**37.61 %**	40.68 %	0.296	**15.70 %**	4.25

Means over 12 seeds, 48 000 episodes per configuration.

The isolated critic without anti-drift reward achieves higher CCR and RCI than the full architecture at higher cost. Full fixed achieves the lowest FSR at the cost of a 15.70 % justified abstention rate. No configuration Pareto-dominates across all four metrics.

The Discussion section should formalise deployment choice as a parameterised loss function: under a loss that penalises false success heavily (security research, legal review, medical decision support), full fixed is preferred; under a loss that penalises missed coverage, critic only dominates.

Venue Guidance

Hard constraint. The ICLR 2027 submission deadline is 25 September 2026 — five days from this document's date. The paper cannot be submitted to ICLR 2027 in its current form. Do not assume a matching workshop exists before the relevant call is published.

1. **Immediate.** Upload to arXiv with a "mechanistic simulation study" descriptor. arXiv is not peer review.
2. **Short term (1–2 months).** Apply C1, C2, I1, I2, I3. Run the pre-registered confirmatory simulation. Update the preprint.
3. **Medium term.** Target workshops that explicitly accept simulation studies or formal analyses of LLM agent reliability. Verify the call before submission.
4. **After LLM validation.** NeurIPS, ICML, or ICLR main tracks. A focused experiment on GAIA [17] or HotpotQA [22] with one open-weight model would materially strengthen suitability for main-track submission.

Summary

DPAD is a well-motivated, formally coherent, and unusually transparent proposal-level paper whose primary limitations are: a critic conditioned on action type unavailable at deployment (C1), and a coupling model in which ρ directly determines failure-indicator correlation rather than emerging from the simulation (C2). After correcting both — together

with the sign error in the latent-variable model (Remark 3) — the paper's simulation-specific empirical validity rises from 3.8 to an estimated 7.0 without any LLM experiment.

P1 and P3 survive C1 unchanged. P2 and T1 require re-derivation if the full-step TPR function differs from the current model. The best-of-N baseline must be split into an oracle upper bound and a verifier-based fair comparison. Calibration of the critic posterior for the P3 test requires a held-out calibration set with a reliability curve and ECE or Brier score, not a fitted score alone.

If C1, C2, I1, I2, and I3 are fully implemented and validated, the paper becomes a strong formal simulation study and a viable candidate for a venue that accepts this class of contribution. The claim that DPAD improves trained LLMs requires the 11 LLM experiments listed in Section 17 of the paper.

Yuval Fradkin — September 20, 2026

References

[1] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

[2] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30.

[3] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv:2110.14168*.

[4] DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill. ISBN 0-07-016242-5.

[5] 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*.

[6] 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 the Association for Computational Linguistics: ACL 2026*, pp. 1975–1991, San Diego, California. *arXiv:2601.22154*.

[7] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In *Proceedings of ICML 2023*, PMLR 202, pp. 10835–10866. `https://proceedings.mlr.press/v202/gao23h.html`

[8] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In *ICLR 2024* (oral). `https://openreview.net/forum?id=VTF8yNQM66`

[9] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.

[10] 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:2608.17776*. Submitted 18 August 2026.

[11] Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Tomasev, N., Balwit, M., Mankowitz, D. J., Perolat, J., & Legg, S. (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`

[12] 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*, pp. 5315–5333. `https://aclanthology.org/2023.acl-long.291/`

[13] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv:2305.19118*.

[14] 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

[15] 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*. https://openreview.net/pdf/a1e5782d4488277b71a1a194ebecc820d981e887.pdf

[16] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442–451. DOI: 10.1016/0005-2795(75)90109-9. PMID: 1180967.

[17] Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2024). GAIA: A benchmark for general AI assistants. In *ICLR 2024*. *arXiv:2311.12983*. https://openreview.net/forum?id=7VPnPnCkYR

[18] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 27730–27744.

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

[20] Uesato, J., Kumar, A., Sifre, L., Hutter, F., Kohli, P., & Bakhtin, A. (2022). Solving math word problems with process- and outcome-based feedback. *arXiv:2211.14275*.

[21] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*. *arXiv:2203.11171*.

[22] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP 2018*, pp. 2369–2380. https://aclanthology.org/D18-1259/

[23] Zhang, B., Zhu, J., Shi, Z., Liu, D., & Tang, R. (2026). AgentForesight: Online auditing for early failure prediction in multi-agent systems. *arXiv:2605.08715* (submitted 9 May 2026). https://arxiv.org/abs/2605.08715