WHEN BETTER GETS WORSE: IMPROVEMENT FIDELITY FOR SELF-IMPROVING AGENTS IN ADAPTIVE WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improving agents increasingly rely on proxy verifiers to choose policy updates, yet deployment can change the world in which those updates are evaluated. An update that looks better to the verifier can therefore become worse after deployment even when the verifier ranks policies well overall. We formalize this gap as *Improvement Fidelity*, which asks whether proxy improvements preserve the sign and ordering of deployment improvements over the updates an improvement process actually proposes. We show that global policy accuracy need not guarantee update fidelity: operator shift and deployment response can create update-level errors, while candidate margins determine whether those errors change the replacement decision. We introduce **PIVOT-KG**, a paired, decision-aware validator that allocates scarce high-fidelity evaluation according to the expected reduction in selection regret per unit cost. Across 90 held-out roots in Leduc, Kuhn, and Melting Pot, proxy and deployment optimal sets are disjoint in 51 cases. In an eight-candidate HighwayEnv stress test, PIVOT-KG reduces mean improvement-selection regret from 0.0435 under the exact Uniform validation rule to 0.0055 at the primary budget. Together, these results show why reliable self-improvement should evaluate proposed improvements in the worlds they induce, while providing a practical rule for allocating scarce deployment evidence when it can affect the replacement decision.[1]

1 INTRODUCTION

Self-improving agents repeatedly propose candidate policies, score them with a verifier, and deploy a replacement. Yet an update that looks better to the verifier can become worse after deployment when the deployed policy changes the world in which it is evaluated. A lane change alters nearby trajectories, and an action in a multi-agent game changes how other agents respond. A verifier can therefore be accurate over policies in aggregate and still choose the wrong local replacement. Figure 1 illustrates this failure mode.

We call this transition-level requirement *Improvement Fidelity*: whether an improvement under the verifier remains an improvement after deployment in the world induced by the update. Because deployment evaluation is often too costly for every candidate, we introduce **PIVOT-KG**, the knowledge-gradient instantiation of PIVOT (Paired Interventional Validation of Optimization Transitions), a paired, decision-aware validator that directs a limited high-fidelity (HF) budget to uncertainty that can change the replacement decision.

Prior work separately studies deployment-induced response, policy-level evaluation or safe improvement, and verifier reliability. We instead study an operator-relative criterion for whether a proposed replacement remains an improvement in the world it induces, together with a decision-aware rule for allocating scarce deployment evidence for that decision.

[1] Our code is available at https://anonymous.4open.science/r/pivot-anonymous-review-2E67/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

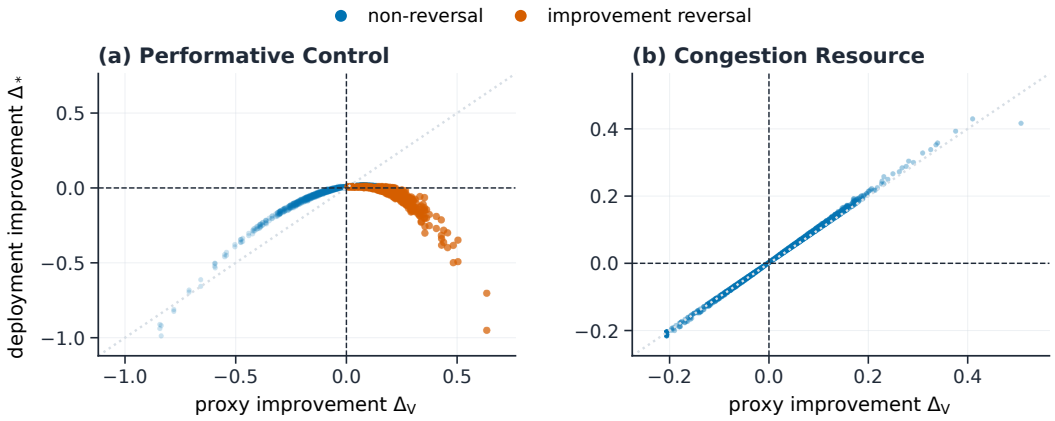

Figure 1: Proxy improvements can reverse after deployment. Each point is a proposed update; lower-right points are proxy-positive but deployment-negative.

Our contributions are:

1. Improvement Fidelity. We formulate reliable self-improvement as an operator-relative, transition-level evaluation problem and introduce Improvement Fidelity, showing why strong global policy evaluation need not imply reliable evaluation of the updates an agent actually proposes.

2. Decision-aware validation. We introduce PIVOT-KG, which allocates a limited HF evaluation budget by its expected reduction in update-selection regret rather than global prediction error.

3. Evidence across adaptive worlds. Across controlled response worlds, multi-agent environments, and reactive driving simulations, we test the failure mechanisms predicted by the theory and identify when targeted HF validation reduces selection regret.

Together, these contributions shift the evaluation of self-improvement from whether policies are evaluated accurately in a proxy world to whether proposed improvements remain improvements after deployment, while providing a principled basis for allocating deployment evidence when it can affect the replacement decision.

2 Related Work

Performative and strategic response. Performative prediction and reinforcement learning study settings in which deployment changes the data distribution, dynamics, rewards, or behavior of other agents (Perdomo et al., 2020; Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2025; Góis et al., 2025; Piliouras & Yu, 2022). These works study how deployment-induced response affects learning, stability, and optimality. We instead ask a local evaluation question: whether a particular proposed replacement remains an improvement under the response world it induces.

Policy evaluation and safe improvement. Off-policy evaluation and safe policy improvement estimate or certify the performance of a candidate relative to a baseline (Jiang & Li, 2016; Thomas et al., 2015; Laroche et al., 2019), while work on learned simulators studies policy evaluation, policy-value gaps, or ranking preservation (Quevedo et al., 2026; Dann et al., 2026). Our focus is the proposed transition itself: a proxy can preserve policy-level accuracy or rankings yet still induce the wrong local replacement decision once deployment changes the evaluation world.

Self-improvement and verification. Recent work has made iterative policy or agent improvement operational (Wang et al., 2026; Zhang et al., 2025) and has documented regressions or failures of verification during self-improvement (Wu et al., 2025; Guo et al., 2026). Related work on reward overoptimization shows that optimizing an imperfect proxy can itself degrade true performance (Gao et al., 2023; Laidlaw et al., 2025). Our failure mode is complementary: the verifier need not be

self-authored, changing, or directly exploited. Even a fixed external verifier can select the wrong update when deployment changes the world in which that update is evaluated.

Adaptive selection and paired evaluation. Knowledge-gradient, ranking-and-selection, and best-arm methods allocate costly observations to decision-relevant alternatives under limited budgets (Chick & Inoue, 2001; Frazier et al., 2009; Kalyanakrishnan et al., 2012; Xie et al., 2016). PIVOT-KG applies this decision-aware allocation principle to proxy–deployment validation, modeling the correction in improvement and valuing HF queries by their effect on the replacement decision.

The gap we study is therefore operator-relative and update-level: whether the specific replacements an improvement process actually proposes preserve the sign and ordering of improvement under the worlds induced by their deployment, and how scarce deployment evidence should be allocated when that uncertainty can change the replacement decision.

3 Improvement Fidelity: Problem Formulation

Proxy and deployment improvement. Let $\mathcal{M}[\pi]$ denote the deployment world induced by policy π, including the environment response to that policy. Let V denote the cheap verifier, with value $J_V(\pi)$, and let $J(\pi; \mathcal{M}[\pi])$ denote the deployment value of π in its induced world. For a proposed replacement $\tau = (\pi, \pi')$, define

$$\Delta_V = J_V(\pi') - J_V(\pi), \qquad \Delta_* = J(\pi'; \mathcal{M}[\pi']) - J(\pi; \mathcal{M}[\pi]). \tag{1}$$

Thus Δ_V is the verifier-estimated improvement and Δ_* is the deployment improvement.

Operator-relative Improvement Fidelity. Let an improvement operator $\mathcal{A}$ induce a transition law $Q_{\mathcal{A}}$ over proposed replacements. For a transition-level loss L, we quantify the corresponding Improvement Fidelity error by

$$\mathcal{E}_{\mathrm{IF}}(V, \mathcal{A}; L) = \mathbb{E}_{\tau \sim Q_{\mathcal{A}}}[L(\Delta_V, \Delta_*)]. \tag{2}$$

The relevant population is the updates the improvement process actually proposes rather than the policy space in general; smaller values indicate greater fidelity.

Fidelity and decision metrics. Improvement Differential Error (IDE) is the absolute-error instance of Eq. (2). We define $\mathrm{IDE} = \mathbb{E}_{Q_{\mathcal{A}}}[|\Delta_V - \Delta_*|]$ for error magnitude, $\mathrm{ISC} = \Pr_{Q_{\mathcal{A}}}(\operatorname{sgn}\Delta_V = \operatorname{sgn}\Delta_* \mid \Delta_V\Delta_* \neq 0)$ for sign agreement after removing ties, and $\mathrm{IRR} = \Pr_{Q_{\mathcal{A}}}(\Delta_* < 0 \mid \Delta_V > 0)$ for the proxy-positive reversal rate. Empirical results report the corresponding sample estimates. Let $C = \{1, \ldots, K\}$ index candidate updates and $C^+ = \{0\} \cup C$ include the no-update action, with $\Delta_{V,0} = \Delta_{*,0} = 0$. For a selected action $\hat{j} \in C^+$, define improvement-selection regret (ISR) as

$$\mathrm{ISR}(\hat{j}) = \max_{j \in C^+} \Delta_{*,j} - \Delta_{*,\hat{j}}. \tag{3}$$

IDE, ISC, and IRR diagnose transition-level fidelity; ISR measures its consequence for the replacement decision.

Deployment response levels. To separate sources of deployment response, we consider the following levels. In the strategic case, π denotes the focal agent i's policy and $R_{-i}(\pi)$ the response of the other agents:

$$\begin{aligned}
\Delta_{\mathrm{direct}} &= J(\pi'; \mathcal{M}[\pi]) - J(\pi; \mathcal{M}[\pi]), \\
\Delta_{\mathrm{actor}} &= J(\pi'; \mathcal{M}[\pi']) - J(\pi; \mathcal{M}[\pi]), \\
\Delta_{\mathrm{strategic}} &= J_i(\pi', R_{-i}(\pi')) - J_i(\pi, R_{-i}(\pi)).
\end{aligned}$$

The direct level holds the incumbent world fixed, the actor level allows the environment to respond to the candidate, and the strategic level additionally allows other agents to adapt through a response map R_{-i}. Here J_i denotes realized value after both environment and other-agent response; best response is one special case. In settings without other-agent adaptation, $\Delta_{\mathrm{actor}} = \Delta_*$.

4 PROPERTIES OF IMPROVEMENT FIDELITY

The results below establish when policy-level evaluation controls update-level fidelity, how that connection can fail, and when the resulting error changes a replacement decision.

From policy fidelity to update fidelity.

Proposition 1 (Uniform value fidelity is sufficient). *If* $\sup_\pi |J_V(\pi) - J(\pi; \mathcal{M}[\pi])| \le \varepsilon$, *then* $|\Delta_V - \Delta_*| \le 2\varepsilon$. *The sign is preserved whenever* $|\Delta_*| > 2\varepsilon$.

Proof. $\Delta_V - \Delta_* = [J_V(\pi') - J(\pi'; \mathcal{M}[\pi'])] - [J_V(\pi) - J(\pi; \mathcal{M}[\pi])]$; the triangle inequality gives the bound and the sign statement. □

Proposition 2 (Global policy accuracy need not imply update fidelity). *For every* $\varepsilon > 0$, *there exist a finite policy family, a verifier, a policy distribution, and an improvement operator with policy mean absolute error (MAE) and Spearman rank error below* ε, *but* $\mathrm{IRR} = 1$ *and* $\mathrm{ISC} = 0$ *under the operator-induced transition law.*

Proof. Let $J(\pi_k) = k/(n-1)$ for $k = 0, \dots, n-1$ under the uniform policy distribution, and let J_V swap the values of adjacent policies π_k, π_{k+1} for any fixed $k \in \{0, \dots, n-2\}$. Then $\mathrm{MAE} = 2/[n(n-1)]$ and $1 - \rho_S = 12/[n(n^2-1)]$, both tending to zero. If the operator always proposes (π_{k+1}, π_k), then $\Delta_V > 0$ while $\Delta_* < 0$, giving $\mathrm{IRR} = 1$ and $\mathrm{ISC} = 0$. □

Operator shift and deployment response.

Proposition 3 (Operator shift bound). *Let* P *be a calibration transition law and* $Q_{\mathcal{A}} \ll P$ *with* $\chi^2(Q_{\mathcal{A}} \| P) < \infty$. *For* $w = dQ_{\mathcal{A}}/dP$ *and* $\ell(\tau) = L(\Delta_V, \Delta_*)$ *with finite variance under* P,

$$|\mathbb{E}_{Q_{\mathcal{A}}}[\ell] - \mathbb{E}_P[\ell]| \le \sqrt{\mathrm{Var}_P(\ell)\, \chi^2(Q_{\mathcal{A}} \| P)}. \tag{4}$$

Proof. Since $\mathbb{E}_P[w] = 1$, $\mathbb{E}_{Q_{\mathcal{A}}}[\ell] - \mathbb{E}_P[\ell] = \mathbb{E}_P[(w-1)(\ell - \mathbb{E}_P[\ell])]$. Cauchy–Schwarz and $\mathbb{E}_P[(w-1)^2] = \chi^2(Q_{\mathcal{A}} \| P)$ give the result. □

Deployment response is a distinct source of error: $\Delta_{\mathrm{actor}} - \Delta_{\mathrm{direct}} = J(\pi'; \mathcal{M}[\pi']) - J(\pi'; \mathcal{M}[\pi])$. The actor–direct gap therefore isolates the change in candidate value caused by the world it induces; the strategic level additionally captures other-agent response through R_{-i}.

When fidelity errors change decisions.

Proposition 4 (Decision preservation). *Let* $\varepsilon_C = \max_{j \in C^+} |\Delta_{V,j} - \Delta_{*,j}|$, $j^* \in \arg\max_{j \in C^+} \Delta_{*,j}$, *and* $\hat{j}_V \in \arg\max_{j \in C^+} \Delta_{V,j}$. *If* j^* *is unique, let* $m = \Delta_{*,j^*} - \max_{k \ne j^*} \Delta_{*,k}$. *Then*

$$\mathrm{ISR}(\hat{j}_V) \le 2\varepsilon_C, \qquad m > 2\varepsilon_C \implies \hat{j}_V = j^*. \tag{5}$$

Proof. Optimality of $\hat{j}_V$ gives $\Delta_{*,j^*} \le \Delta_{V,j^*} + \varepsilon_C \le \Delta_{V,\hat{j}_V} + \varepsilon_C \le \Delta_{*,\hat{j}_V} + 2\varepsilon_C$. The same error bound preserves the top-1 ordering when $m > 2\varepsilon_C$. □

With matched randomness, positive incumbent–candidate covariance reduces the variance of their paired difference relative to independent evaluation with the same marginal variances. Uniform policy fidelity is sufficient to control update error, but global average accuracy need not protect operator-induced updates. Operator shift and deployment response create distinct sources of update-level error, and these errors change replacement decisions only relative to candidate margins. This motivates directing scarce HF evidence to uncertainty that can change the selected update.

5 PIVOT-KG: DECISION-AWARE VALIDATION

PIVOT is a paired validation framework for candidate transitions; PIVOT-KG is its decision-aware instantiation using the knowledge-gradient acquisition rule below. We therefore allocate expensive high-fidelity (HF) evaluation to uncertainty that can change the replacement decision. Given an incumbent and K candidate updates, the cheap verifier provides proxy improvements $\{\Delta_{V,j}\}_{j=1}^{K}$. PIVOT-KG models the proxy-to-deployment correction $G_j = \Delta_{*,j} - \Delta_{V,j}$.

For tractable sequential updating, we use a calibrated working Gaussian correction model over candidate-wise vectors $G, \Delta_V, \Delta_* \in \mathbb{R}^K$:

$$G \sim \mathcal{N}(\mu, \Sigma_G), \qquad \Delta_* \mid \Delta_V \sim \mathcal{N}(\Delta_V + \mu, \Sigma_G).$$

The prior parameters and observation variances are estimated from a disjoint calibration set of paired proxy/HF transitions, so each held-out decision root starts from this calibrated belief without using its HF outcomes.

A query of candidate j produces a prescribed paired HF evaluation $Y_j = \Delta_{*,j} + \eta_j$ with $\eta_j \sim \mathcal{N}(0, s_j^2)$, where s_j^2 is estimated and regularized from calibration data. Gaussian conditioning uses a working likelihood with conditionally independent query noise. The incumbent and candidate are evaluated under matched initial state, exogenous randomness, and opponent initialization. Each query updates the joint posterior by Gaussian conditioning; non-diagonal Σ_G lets evidence update correlated unqueried candidates. In the fixed-budget procedure, each candidate is queried at most once.

Given queried data D, let $a(D)$ denote the posterior replacement decision and $\mathcal{R}(D)$ its Bayes selection regret. Using C^+ defined above,

$$a(D) = \arg\max_{j \in C^+} \mathbb{E}[\Delta_{*,j} \mid D],$$
$$\mathcal{R}(D) = \mathbb{E}\left[\max_{j \in C^+} \Delta_{*,j} - \Delta_{*,a(D)} \,\middle|\, D\right].$$

For a hypothetical HF observation Y_j, define its expected value of sample information (EVSI) and cost-normalized acquisition score as

$$\begin{aligned} \mathrm{EVSI}_j &= \mathcal{R}(D) - \mathbb{E}_{Y_j \mid D}\left[\mathcal{R}\big(D \cup \{(j, Y_j)\}\big)\right], \\ A_j &= \frac{\mathrm{EVSI}_j}{c_j}. \end{aligned} \tag{6}$$

PIVOT-KG therefore queries the feasible candidate with the largest expected reduction in selection regret per unit HF cost. Implementations use Monte Carlo or analytic Gaussian integration for these posterior expectations.

Algorithm 1 PIVOT-KG under a fixed high-fidelity budget

Require: Incumbent π; candidates $\{\pi'_j\}_{j=1}^{K}$; calibrated model $(\mu, \Sigma_G, \{s_j^2\}_{j=1}^{K})$; query costs $\{c_j\}_{j=1}^{K}$; HF budget B
Ensure: Selected candidate or incumbent
1: Evaluate candidates to obtain $\{\Delta_{V,j}\}_{j=1}^{K}$
2: Initialize the calibrated posterior, $D \leftarrow \emptyset$, $S \leftarrow \emptyset$, and $b \leftarrow B$
3: **while** there exists $j \in C \setminus S$ with $c_j \le b$ **do**
4: Compute A_j for each feasible $j \in C \setminus S$
5: $j_q \leftarrow \arg\max_{j \in C \setminus S:\, c_j \le b} A_j$
6: Obtain paired HF observation Y_{j_q} under matched randomness
7: $D \leftarrow D \cup \{(j_q, Y_{j_q})\}$ and condition the posterior
8: $S \leftarrow S \cup \{j_q\}$ and $b \leftarrow b - c_{j_q}$
9: **end while**
10: **return** $a(D)$

The fixed-budget rule is our primary procedure; posterior-based early stopping is a secondary variant.

6 EXPERIMENTS

The experiments follow the progression from update-level fidelity failure to its decision consequences and budgeted validation. We first isolate operator shift and deployment response, then test when the resulting errors alter the replacement decision, and finally evaluate whether PIVOT-KG can reduce selection regret with limited high-fidelity (HF) evidence. We conclude with a reactive driving stress test.

6.1 EXPERIMENTAL PROTOCOL

Controlled response studies use Performative Control and Congestion Resource to isolate operator shift and deployment response; the pooled operator-shift contrast additionally includes a frozen MPE2 (Lowe et al., 2017) condition described in Appendix A. Multi-agent evaluations use Leduc and Kuhn poker through OpenSpiel (Lanctot et al., 2020) and the Melting Pot suite (Leibo et al., 2021). Controlled studies use paired common-random-number rollouts with 30 independent seeds, and the held-out multi-agent study contains 30 roots in each environment. Each held-out decision root has a fixed panel of candidate policy updates generated by its improvement operator; selection also permits the incumbent no-update option. Held-out comparisons use paired root-bootstrap confidence intervals (CIs) conditional on the fitted calibration model; the controlled-study aggregation is specified in Appendix A. An HF query is one prescribed paired candidate evaluation under matched randomness. Primary comparisons use PIVOT-KG and Uniform; controlled batch diagnostics additionally include Proxy Only, value-of-information (VOI) and lower and upper confidence bound (LUCB) heuristics, and an all-HF reference. HighwayEnv (Leurent, 2018) is the primary reactive driving stress test. A second Leduc cohort and a secondary MetaDrive stress test are reported in the appendix. Bootstrap intervals resample the corresponding inferential unit; further reproducibility details are provided in the supplement.

6.2 IMPROVEMENT FIDELITY UNDER OPERATOR SHIFT AND DEPLOYMENT RESPONSE

We first test the two mechanisms identified by the theory. Operator shift changes which proposed transitions matter, whereas deployment response changes the world in which a candidate is evaluated. In the operator-shift study, the pooled high-minus-low contrast in operator-relative IDE is 0.2627 [0.1684, 0.3763] (descriptive cell-bootstrap interval, 30 seed values; Appendix A). Local update error and reversals increase with the designed shift intensity, while global policy rank is a fixed calibration reference (Figure 2). Thus global policy accuracy can mask failure on the updates an improvement operator actually proposes.

Deployment response provides a distinct source of discrepancy. In the strategic fixture, the strategic-minus-actor effect is -0.0240 [-0.0249, -0.0231], with family-balanced strategic reversal rate 0.9495.

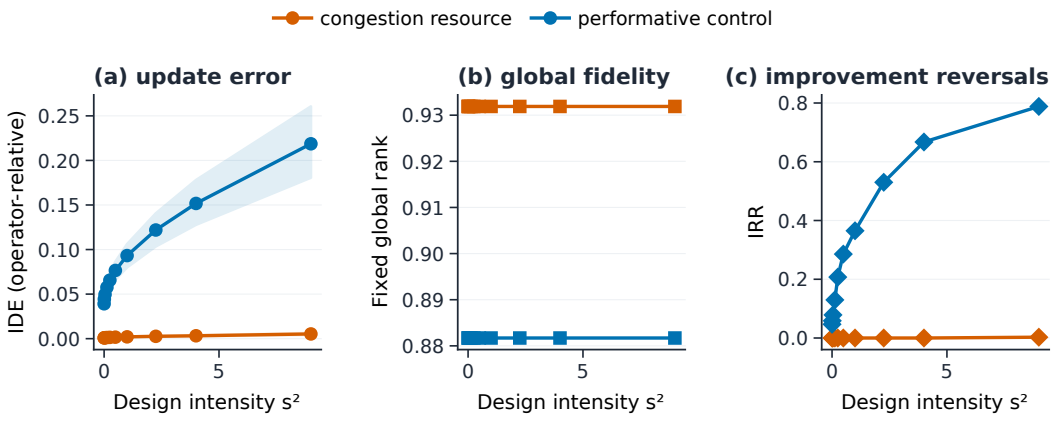

Figure 2: Operator shift exposes update-level failure. The horizontal axis is the design intensity s^2, not a measured χ^2 divergence. Global Spearman rank is a fixed calibration reference.

A second Leduc confirmation cohort shows the same mechanism along the response horizon: the proxy-positive reversal rate rises from 2.5% under one-step response to 51.7% under eight-step adaptation (Appendix A). The appendix also compares response-disabled proxy, actor, and strategic evaluations across opponents. Together, these results identify two distinct sources of update-level failure: operator shift can concentrate error on the transitions proposed by the improvement process, while deployment response can sharply alter the value and sign of a candidate after deployment.

6.3 Decision Relevance and High-Fidelity Resolvability

Proposition 4 shows that update error changes a replacement decision only when it is large relative to the candidate margin. For each held-out root, we compare the deployment winner with its competitors using a winner-relative correction-to-margin ratio; values above one indicate disjoint proxy and deployment optimal sets, with one as the tie boundary. Appendix B gives the exact definition.

Across 90 held-out roots, proxy and deployment optimal sets are disjoint in 51 cases. Figure 3 shows that disagreement alone does not determine validation value: what matters is whether the fidelity error is large enough to threaten the replacement decision and whether available HF evidence can resolve the resulting margin. At the primary long budget, the PIVOT query set intersects the ex-post decision-critical pair in 0.97 [0.92, 1.00] of roots, while the mean model-assumed queried observation standard deviation (SD) is 13.99 [7.97, 22.90] times the proxy top-2 margin. This regularized-model diagnostic does not estimate HF observation noise directly; targeting alone need not resolve ordering.

The held-out analysis isolates the mechanism behind useful validation: fidelity error must threaten the replacement decision, and the available HF evidence must be informative enough to resolve that uncertainty.

6.4 PIVOT-KG under Limited High-Fidelity Budgets

Table 1 reports the primary held-out fixed-budget comparison. PIVOT-KG substantially lowers ISR in Leduc, the contrast is small in Kuhn, and the primary Melting Pot comparison is unresolved. This heterogeneity is consistent with the preceding mechanism analysis: deployment response alone does not determine the value of targeted validation. Secondary Melting Pot conditions are reported in Appendix B. For these cohorts, PIVOT-KG and Uniform share the calibrated posterior, observation-update rule, and final selection rule, differing only in query allocation; related component diagnostics are reported in Appendix B.

Table 1: Primary held-out fixed-budget comparisons (30 paired roots per row). Positive Uniform − PIVOT means lower ISR for PIVOT-KG; brackets give 95% root-bootstrap CIs.

World / response	HF queries	Uniform ISR	PIVOT-KG ISR	Uniform − PIVOT [95% CI]
Leduc, long	2	0.046	0.009	0.036 [0.014, 0.064]
Kuhn, long	2	0.058	0.056	0.002 [0.000, 0.006]
Melting Pot, long	2	1.742	0.836	0.906 [-0.784, 3.401]

We also retain controlled batch-selector diagnostics, distinct from Algorithm 1. Their pooled candidate-only cumulative improvement-selection regret (CISR) contrast for Proxy-Only minus PIVOT-batch under the legacy forced-replacement setting is 1.4883 [1.0618, 1.9105] over 40 rounds; Figure 4 compares the single-decision batch selectors at matched query counts. A second Leduc confirmation cohort yields a prespecified PIVOT-KG versus expected-Uniform selected-gain contrast of +0.02989 using a repeated-query, sample-size variant. Its component comparisons, the legacy closed-loop target, stopping behavior, and numerical diagnostics are specified in Appendix B.

Together, these results support the central prediction of our analysis: decision-aware validation is most useful when the remaining uncertainty is both consequential for the replacement decision and resolvable by the available HF evidence.

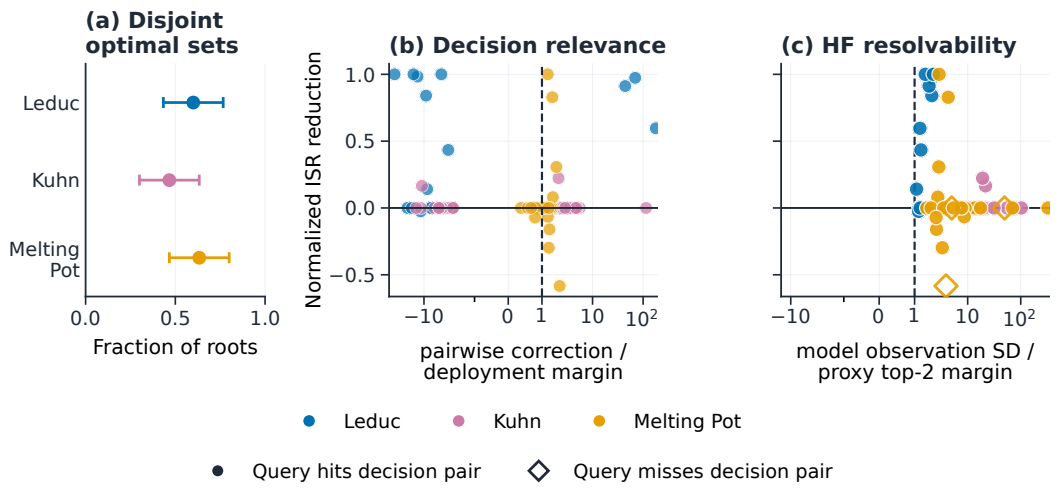

Figure 3: Decision relevance and HF resolvability. (a) Disjoint optimal sets. (b) Winner-relative correction-to-margin ratio versus normalized ISR reduction. (c) Model-assumed observation SD relative to the proxy top-2 margin versus normalized ISR reduction; filled markers hit the ex-post decision-critical pair.

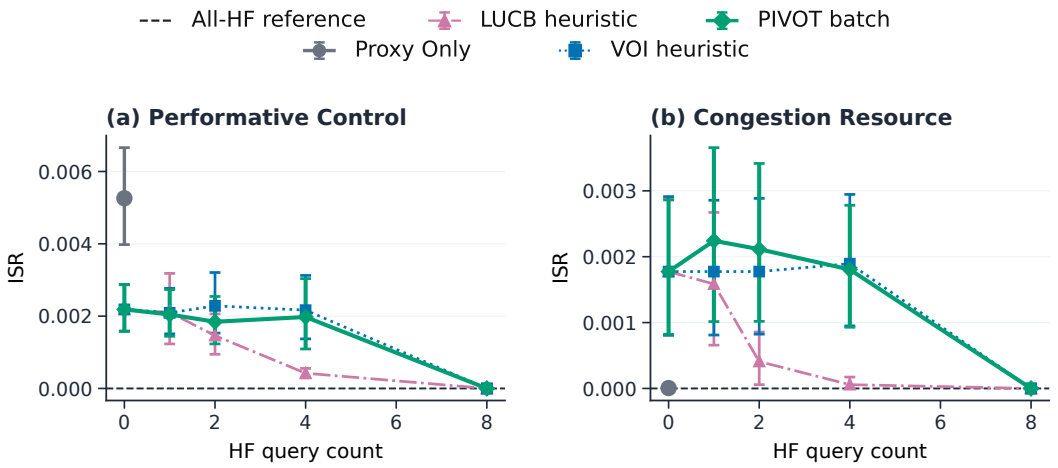

Figure 4: Legacy batch-selector diagnostics at matched HF query counts in (a) Performative Control and (b) Congestion Resource. Lower single-decision ISR is better; the all-HF reference uses only fully queried rows. Intervals bootstrap seeds.

6.5 Reactive Driving Stress Test

We finally test PIVOT-KG in a reactive driving simulation where HF evaluation is scarce relative to the candidate set. The HighwayEnv redesign uses eight scheduled candidates formed by crossing four non-idle actions with early and late half-horizon windows, fits the correction model on 40 calibration roots, and evaluates 120 disjoint test roots. At the primary budget $B = 4$, PIVOT-KG reduces ISR relative to the exact Uniform validation rule by 0.0380 $[0.0300, 0.0474]$. The Uniform rule averages exactly over all feasible queried subsets, and each PIVOT query corresponds to a paired actor rollout in the reactive simulator.

The HighwayEnv result shows the same decision-aware validation principle in a reactive driving candidate-allocation problem. Earlier HighwayEnv cohorts were unresolved at their primary budgets; these results, the secondary $B = 2$ redesign result, and a secondary MetaDrive (Li et al., 2023) stress test are reported in Appendix C.

7 CONCLUSION

This paper reframes reliable self-improvement as an operator-relative, update-level evaluation problem. Improvement Fidelity asks whether a proposed improvement under a proxy remains an improvement after deployment in the world that the update induces. Our analysis shows why strong policy-level evaluation is not sufficient: operator shift and deployment response can create update-level errors, while candidate margins determine whether those errors change the replacement decision. PIVOT-KG turns this structure into a decision-aware validation rule that allocates scarce high-fidelity evidence toward uncertainty that can affect the selected update. Across controlled, multi-agent, and reactive driving simulations, our experiments support this view and show when targeted validation can reduce selection regret.

The current study has three main limitations. First, the direct, actor, and strategic response models capture several important forms of adaptation but do not cover all deployment dynamics. Second, PIVOT-KG relies on a calibrated correction model and sufficiently informative HF observations; when evidence is too noisy or budgets are too small, decision-critical uncertainty can remain unresolved. Reported held-out intervals condition on the fitted calibration model and do not propagate calibration-set sampling uncertainty. Third, our driving-simulator evidence has its clearest allocation result in HighwayEnv and a more mixed MetaDrive stress test.

These results suggest several directions for future work, including richer response and correction models, repeated self-improvement under co-adaptation, and intervention-based HF evaluation in more realistic deployment systems. More broadly, as agents increasingly modify their own policies, the central evaluation question is not only whether a candidate scores better under a verifier, but whether replacing the incumbent with that candidate remains an improvement in the world the replacement creates. When better can become worse after deployment, reliable self-improvement requires fidelity of improvements, not merely accuracy of policy evaluations—a principle that provides both a theoretical criterion for update reliability and a practical basis for allocating scarce deployment evidence when it can affect the replacement decision.

AI USE STATEMENT

The research ideas, problem formulation, methodology, theoretical development, experimental design, analysis, and initial manuscript draft were developed and written by the authors. Generative AI tools were used only to assist with limited portions of code implementation, scientific consistency checks, and final-stage language and presentation polishing. AI-assisted suggestions and edits remain subject to verification by the authors. The authors take full responsibility for the research, experimental results, and final manuscript.

ETHICS STATEMENT

This study uses controlled virtual environments and publicly available benchmark environments and data. It involves no human subjects, personal data, live trading, or real-world autonomous-system deployment. The experiments are conducted entirely in simulation and do not directly affect users or physical systems.

REPRODUCIBILITY STATEMENT

The accompanying supplementary material includes an anonymized code archive, cohort-specific configurations, figure provenance, and reproduction instructions for the reported results, distinguishing legacy batch diagnostics from the held-out and driving implementations. The anonymized code and reproducibility materials are also available at `https://anonymous.4open.science/r/pivot-anonymous-review-2E67/`.

REFERENCES

Stephen E. Chick and Koichiro Inoue. New procedures to select the best simulated system using common random numbers. *Management Science*, 47(8):1133–1149, 2001. doi: 10.1287/mnsc.47.8.1133.10229.

Christoph Dann, Yishay Mansour, and Mehryar Mohri. Theoretical foundations and effective algorithms for policy-aware simulator learning. *arXiv preprint arXiv:2605.29032*, 2026. URL https://arxiv.org/abs/2605.29032.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009. doi: 10.1287/ijoc.1080.0314.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

António Góis, Mehrnaz Mofakhami, Fernando P. Santos, Gauthier Gidel, and Simon Lacoste-Julien. Performative prediction on games and mechanism design. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1855–1863. PMLR, 2025. URL https://proceedings.mlr.press/v258/gois25a.html.

Diandian Guo, Cong Cao, Fangfang Yuan, Yingqi Wang, Yueshan Wang, and Dakui Wang. Self-authored verification is unreliable in heuristic self-improving agents. *arXiv preprint arXiv:2607.24300*, 2026. URL https://arxiv.org/abs/2607.24300.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 652–661, 2016. URL https://arxiv.org/abs/1511.03722.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 655–662, 2012.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dbe2cfe4767f3255160b73a36ae3162e-Abstract-Conference.html.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2020. URL https://arxiv.org/abs/1908.09453.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Joel Z. Leibo, Edgar A. Dueñez-Guzman, Alexander Vezhnevets, John Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6187–6199. PMLR, 2021. URL https://proceedings.mlr.press/v139/leibo21a.html.

Edouard Leurent. An environment for autonomous driving decision-making. https://github.com/Farama-Foundation/HighwayEnv, 2018.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023. doi: 10.1109/TPAMI.2022.3190471.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.02275.

Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear markov decision process. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3232–3240. PMLR, 2025. URL https://proceedings.mlr.press/v258/mandal25a.html.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23642–23680. PMLR, 2023. URL https://proceedings.mlr.press/v202/mandal23a.html.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv preprint arXiv:2201.10483*, 2022. URL https://arxiv.org/abs/2201.10483.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7f5e909ac0324db03506b380c695ffaf-Abstract-Conference.html.

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 3041–3075. PMLR, 2024. URL https://proceedings.mlr.press/v244/rank24a.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Zhilin Wang, Han Song, Runzhe Zhan, Jusen Du, Jiacheng Chen, Tianle Li, Qingyu Yin, Yulun Wu, Zhennan Shen, Tong Zhu, Yanshu Li, Guanjie Chen, Derek F. Wong, Yafu Li, Yu Cheng, and Yang Yang. EvoPolicyGym: Evaluating autonomous policy evolution in interactive environments. *arXiv preprint arXiv:2607.02440*, 2026. URL https://arxiv.org/abs/2607.02440.

Ting Wu, Xuefeng Li, and Pengfei Liu. Progress or regress? self-improvement reversal in post-training. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1fa0c4e5a7e189729230d018b229abc7-Abstract-Conference.html.

Jing Xie, Peter I. Frazier, and Stephen E. Chick. Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research*, 64(2):542–559, 2016. doi: 10.1287/opre.2016.1480.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025. URL https://arxiv.org/abs/2505.22954.

A Response Mechanism Analyses

The operator contrast pools Performative Control, Congestion Resource, and frozen Multi-Agent Particle Environment 2 (MPE2) (Lowe et al., 2017) with weights $0.4/0.4/0.2$; Figure 2 shows the first two. Its interval resamples 450 condition/seed cells sharing 30 seeds; figure bands resample six condition means. Congestion strengths are duplicated; global rank is a fixed zero-shift evolutionary-policy reference. These are descriptive shift diagnostics, not measured χ^2 divergence. Figure 5's proxy disables response, unlike the fixed-incumbent-world Δ_{direct}. Strategic reversal averages the three adaptive families.

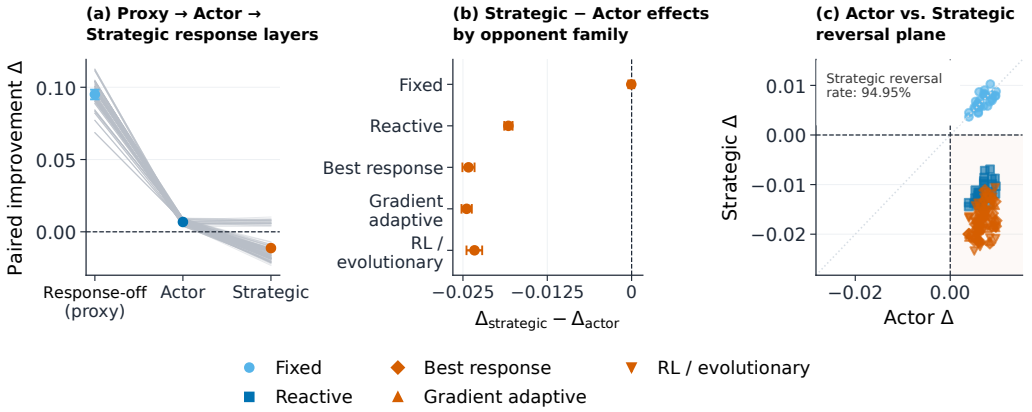

Figure 5: Response mechanisms. (a) Response-disabled proxy, actor, and strategic evaluations. (b) Strategic-minus-actor effects by opponent family. (c) Actor versus strategic improvement. Intervals use 30 matched seeds.

A separate Leduc confirmation cohort tests the same mechanism along the response horizon. Among proxy-positive candidates, the mean root-level improvement-reversal rate rises from 2.5% under one-step response to 51.7% under eight-step adaptation (Figure 6). This supports the main-text conclusion that deployment response can amplify update-level failure rather than merely shift policy values uniformly.

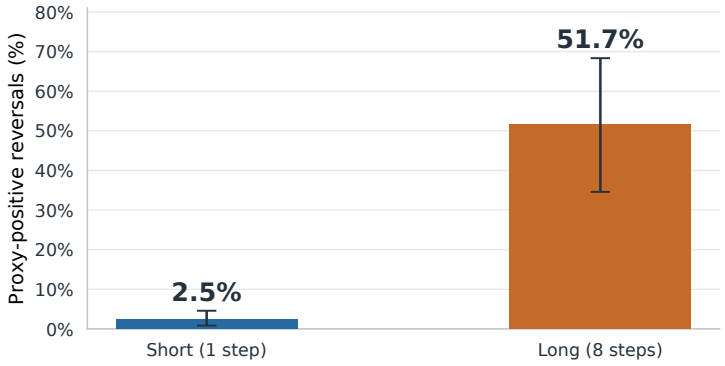

Figure 6: Leduc response-horizon confirmation. Reversal rates are computed among proxy-positive updates within each root, where a reversal means negative deployment gain. Bars show means across 30 roots under one-step and eight-step adaptation, with 95% root-bootstrap CIs.

B EXTENDED PIVOT-KG EVALUATION

In the calibrated poker, Melting Pot, and driving evaluations, candidate coordinates follow the same fixed update slots across calibration and test roots.

Decision diagnostics. Let $T = \arg\max_{C^+} \Delta_*$ and $b \in \arg\max_{j \in T} \Delta_{V,j}$. The ratio is $\max_{k \notin T}(G_b - G_k)/(\Delta_{*,b} - \Delta_{*,k})$, with $G = \Delta_* - \Delta_V$ and tie tolerance 10^{-12}. The 51 disagreements are disjoint optimal sets; 28 roots have tied optima. Above one indicates disjoint optimal sets; one is the tie boundary. The query-hit pair uses proxy/deployment top candidates within C (stable index ties); top-2 margins exclude no-update. Normalization divides Uniform minus PIVOT ISR by the deployment range over C^+. The noise ratio uses mean queried $\sqrt{R_j}$, a margin floor of 10^{-12}, and poker's model-variance floor 0.25; it is model-assumed, not directly measured observation noise.

At the primary long budget, the decision-critical query-hit rates are 1.00 in Leduc, 1.00 in Kuhn, and 0.90 in Melting Pot. The descriptive normalized gain is 0.113 [0.055, 0.180] among the 87 hit roots and -0.195 [-0.584, 0.000] among the three misses. These diagnostics separate successful targeting from successful resolution.

Secondary and confirmation analyses. The first cohorts use 12 calibration/30 test roots and one frozen Uniform draw. Poker's nine queryable entries include an incumbent copy; Melting Pot uses eight specialist mixtures on one repeated stag-hunt substrate. Full candidate/cost protocols accompany the supplement. Designated allocation interactions are $0.0365\,[0.0145, 0.0637]$ (Leduc), $0.0020\,[0, 0.0058]$ (Kuhn), and $1.4815\,[-0.3930, 4.0956]$ (Melting Pot); only Leduc meets the joint primary/interaction criterion.

Table 2 retains the secondary comparisons. The second Leduc cohort was designed after the first-cohort results, with fresh confirmation roots. This cohort is a mechanism confirmation study rather than a direct replication of the fixed-budget benchmark. Its PIVOT-KG variant queries (j, n) pairs with $n \in \{512, 2048, 4096\}$, permits repeated candidates, and charges $2n$ hands; it extends the once-per-candidate Algorithm 1. At 16,384 hands its prespecified contrast against exact-subset expected Uniform is +0.02989 under long adaptation. Figure 7 retains the component comparisons; its final row is $(\mathrm{PIVOT} - \mathrm{Uniform})_{\mathrm{long}} - (\mathrm{PIVOT} - \mathrm{Uniform})_{\mathrm{short}}$, not a direct long-minus-short policy-gain contrast.

Table 2: Secondary held-out comparisons. Positive Uniform − PIVOT means lower ISR for PIVOT-KG; brackets give 95% root-bootstrap CIs.

World / response	HF queries	Uniform ISR	PIVOT-KG ISR	Uniform − PIVOT [95% CI]
Melting Pot, long	1	5.283	1.859	3.423 [0.390, 6.944]
Melting Pot, short	5	1.035	1.610	-0.575 [-1.042, -0.180]

The secondary stopping rule reduces mean HF use from 16,384 to 12,083.2 hands, a 26.25% reduction, while producing the same selected gain as the full-budget rule on all 30 observed roots. The paired difference in HF use is $-4300.8\ [-6007.5, -2662.4]$ hands. Thus the stopping variant can save validation queries in this cohort without changing the observed replacement decision.

Closed-loop and numerical diagnostics. Figures 4, 8, and 9 use legacy batch selection, not Algorithm 1: no within-batch reacquisition; observed values replace queried estimates; unqueried estimates can omit later observations. The closed loop forces replacement and reports $\mathrm{CISR}_C = \sum_t [\max_{j \in C_t} \Delta_{*,j} - \Delta_{*,\hat{j}_t}]$ without no-update; negative improvement with zero regret is then possible. Figure 4 counts queries and uses VOI/LUCB heuristics. Figure 9 compares Monte Carlo draws on 24 Congestion test groups ($K = 4$): all costs scale equally, query count stays two, and seeds vary. With common draws the ranking is invariant, so this does not identify cost misspecification.

Additional response-mixture settings yield positive, null, and negative allocation effects, consistent with the main-text conclusion that deployment response alone does not determine validation value. Full method-by-method and response-mixture diagnostics are provided in the supplement.

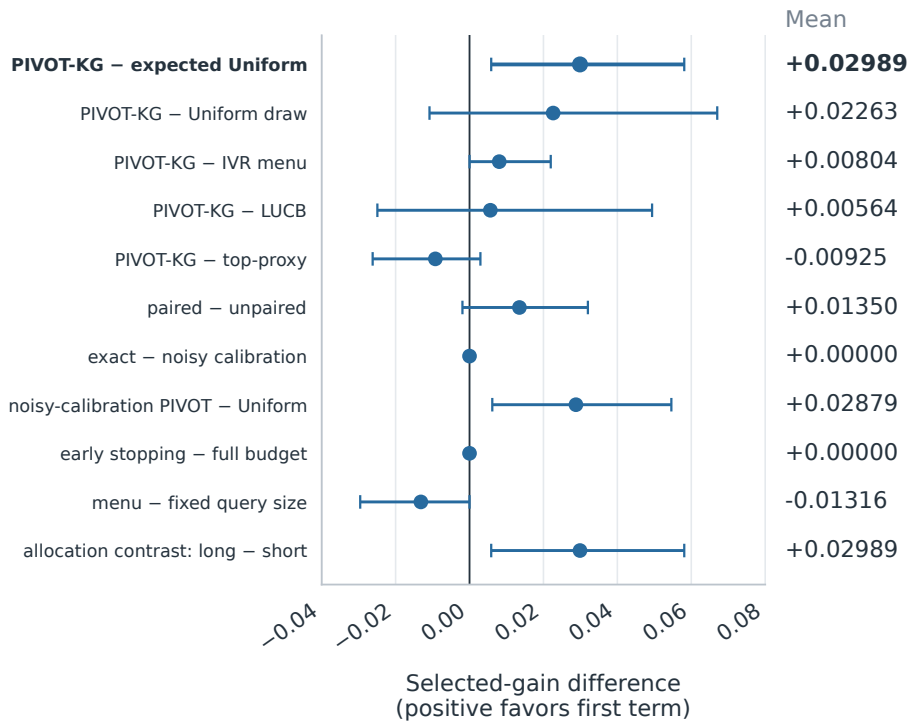

Figure 7: Leduc confirmation-cohort component contrasts under long response (30 roots; primary budget 16,384 HF hands). Positive values favor the first term. The expected-Uniform comparison is prespecified primary; all others are secondary with unadjusted 95% CIs.

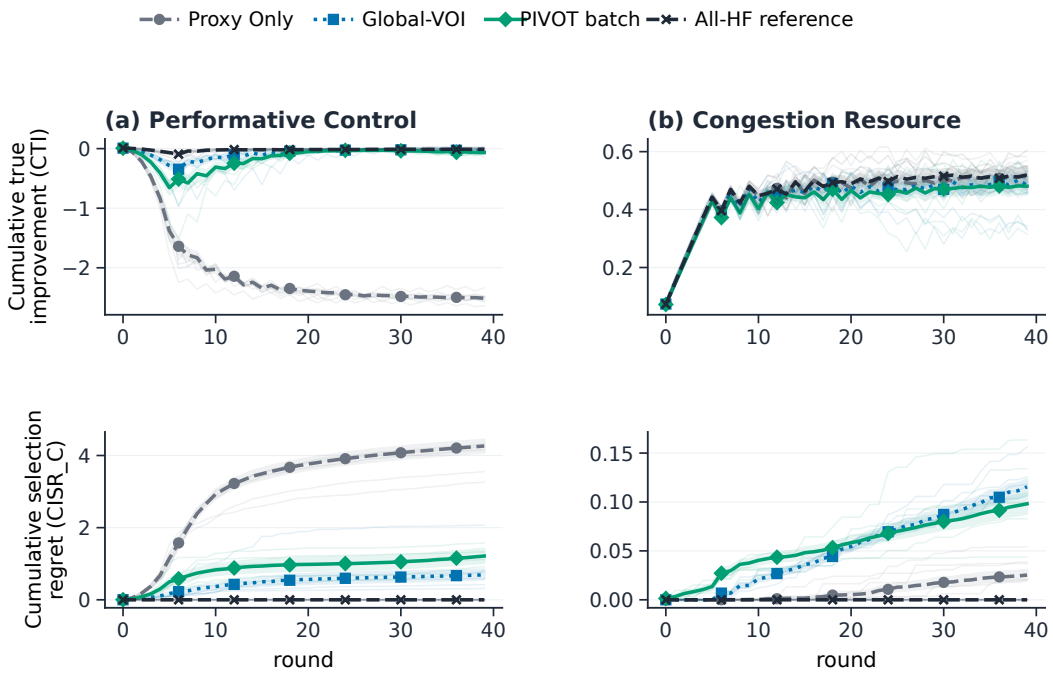

Figure 8: Legacy forced-replacement trajectories. Top: cumulative true improvement; bottom: candidate-only CISR_C, excluding no-update. Thin curves are seeds; markers summarize batch decision rules.

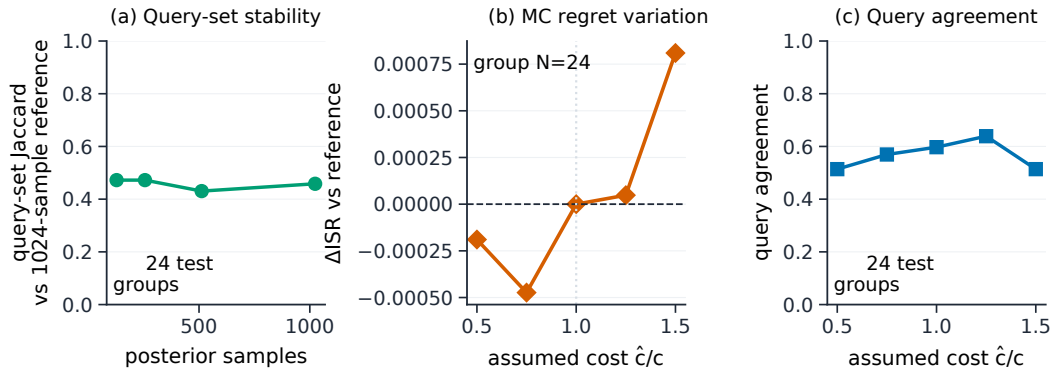

Figure 9: Legacy batch-selector Monte Carlo diagnostics on 24 test groups. Panels show query agreement across sample counts and regret/query variation under common cost scaling with different random draws; this is not an identified cost-misspecification effect.

C Reactive Driving Stress Tests

HighwayEnv. The original and fresh-seed replication cohorts each use 20 calibration and 60 test roots, with no overlap between cohorts. Both use the same five-action panel; the primary budget is $B = 2$ and $B = 4$ is secondary. Intervals use paired root bootstrap and condition on the fitted calibration model. A technical smoke run exposed two replication roots; they are retained without subsequent tuning or exclusion. Uniform retains proxy estimates for unqueried candidates, whereas PIVOT-KG uses its calibrated posterior, so the comparison evaluates complete selection rules rather than acquisition alone.

The eight-candidate redesign freezes eight scheduled candidates, fits a joint correction model on 40 calibration roots, and evaluates 120 disjoint test roots. Its primary budget is $B = 4$ and $B = 2$ is secondary. Uniform is the exact average over all $\binom{8}{B}$ queried subsets (70 at $B = 4$ and 28 at $B = 2$), while PIVOT-KG applies sequential Gaussian updates. Each query is a paired actor rollout in the reactive simulator. Two test roots were exposed during technical smoke and are retained without retuning or exclusion. Detailed execution and audit records are provided in the supplement.

Table 3: HighwayEnv cohorts. Positive Uniform − PIVOT means lower ISR for PIVOT-KG; brackets give paired 95% bootstrap CIs. Bold budgets mark the primary comparison within each cohort.

Cohort	B	Roots	Uniform ISR	PIVOT-KG ISR	Uniform − PIVOT [95% CI]
Original	1	60	0.0492	0.0718	-0.0226 $[-0.0439, -0.0008]$
Original	**2**	60	0.0422	0.0374	0.0049 $[-0.0088, 0.0199]$
Original	4	60	0.0144	0.0029	0.0115 $[0.0008, 0.0237]$
Replication	**2**	60	0.0254	0.0261	-0.0006 $[-0.0143, 0.0145]$
Replication	4	60	0.0037	0.0004	0.0034 $[-0.0007, 0.0099]$
Redesign (8)	2	120	0.0613	0.0157	0.0456 $[0.0343, 0.0587]$
Redesign (8)	**4**	120	0.0435	0.0055	0.0380 $[0.0300, 0.0474]$

MetaDrive. We additionally evaluate a first MetaDrive (Li et al., 2023) cohort under a long-response condition with 12 profiles and 30 held-out roots. Figure 10 reports selected deployment gain across one, two, and four HF candidate-query budgets. At the prespecified two-query budget, the paired PIVOT-KG–Uniform contrast is $+1.89$ $[-1.29, 5.49]$ (30-root bootstrap, 10,000 draws), leaving the primary comparison unresolved. The $B = 1$ and $B = 4$ comparisons are secondary with unadjusted intervals. We retain MetaDrive as boundary evidence rather than a stable positive result; roots share one fixed bottleneck geometry. Integrated variance reduction (IVR) and LUCB use the matched posterior. The calibrated no-HF and proxy-only baselines incur zero HF cost and appear as the

separate No-HF baselines. Uniform averages 100 allocation draws per root. Positive-query costs include response search and paired evaluation; the supplement retains all eight rules.

Evaluator boundary. Improvement Fidelity defines an evaluation target, not a preferred learner architecture. In an out-of-distribution comparison, a transition-specific learner does not outperform the global learner: the transition-minus-global ISC contrast is -0.2590 [-0.3588, -0.1707]. Full ISC and IDE comparisons are reported in the supplement.

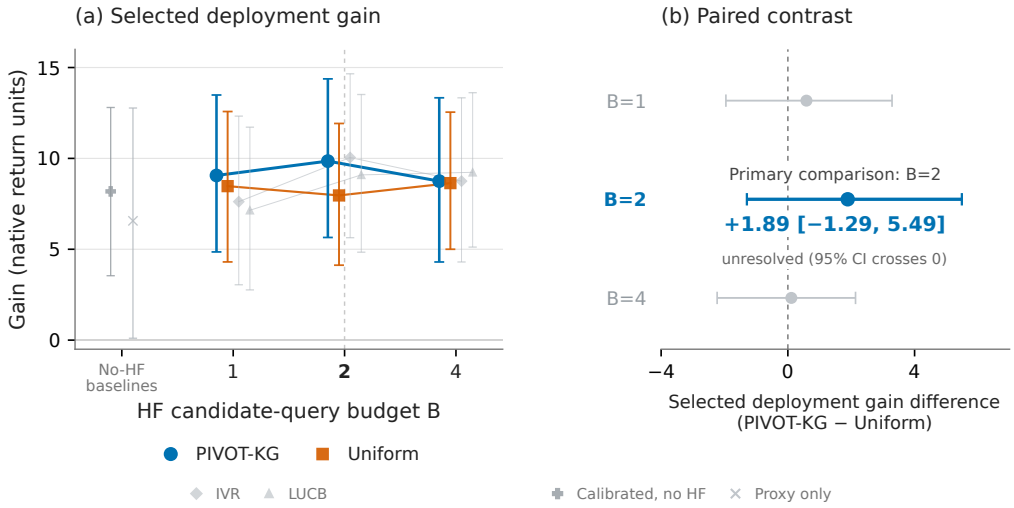

Figure 10: MetaDrive stress test as a boundary case. The prespecified $B = 2$ comparison remains unresolved: $+1.89\,[-1.29, 5.49]$, with the 95% root-bootstrap CI crossing zero. (a) Selected deployment gain (12 profiles, 30 held-out roots; marginal 95% root-bootstrap intervals). (b) Paired PIVOT-KG–Uniform selected deployment gain contrasts; $B = 1, 4$ are secondary.