FROM DIVERGENCE SCORES TO AUDITABLE EVIDENCE: EXACT-TV ANCHORS FOR REWARD-INTERFACE DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Trustworthy paired-rollout audits require an observable behavioral target and a conclusion that survives independent administration. We meet both requirements for reward-aware agents by measuring how a frozen policy's action distribution responds to task-preserving edits of shaping meters, verifier messages, judge feedback, and reward-model outputs. The Paired Interface Response Audit (PIRA) admits edits through ground-truth preservation tests, estimates an intervention-indexed response profile from matched rollouts, and returns a finite-sample rank-conformal Holm decision (PIRA-Holm) together with a learned ranking score (PIRA-L). Three open-weight exact-TV anchors expose the target action-distribution shift directly; frozen reruns and low-overlap external banks then test execution and bank-design invariance. PIRA-Holm attains $0.872/0.868/0.866$ AUROC and PIRA-L attains $0.903\pm.012/0.897\pm.013/0.894\pm.013$ on τ-bench-Exact, SWE-Gym-Exact, and MLE-Bench-Exact. Independent operators reproduce these results within $0.005/0.006/0.006$ AUROC. Three full external handoffs with 8.6%–11.2% semantic overlap reach $0.884\pm.015$, $0.882\pm.014$, and $0.879\pm.014$, retain worst-family AUROC of at least 0.802, and keep Benign-Hard false-positive rates below 5%. These results turn reward-interface sensitivity from a score correlate into an exact-anchored interventional measurement and give paired-rollout audits a reproducible evidentiary bar.

1 INTRODUCTION

Reward-aware agents consume evaluation signals as part of their observation stream. Shaping bonuses steer control policies, unit-test feedback redirects coding agents, and judge or verifier messages influence tool-using language models. Altering one of these channels changes the information available at the next decision, so a frozen policy can exhibit a substantial interface-conditioned action shift. Detecting this dependence before deployment reveals when an agent's behavior is coupled to evaluator-facing details that preserve the underlying task.

The scientific challenge is to identify that action shift as a measured object. Exploit-versus-benign discrimination can mix interface dependence with task difficulty, trajectory style, or classifier capacity. An interventional audit must instead qualify the edit, isolate policy response from environmental randomness, expose the target quantity on systems where it is exactly observable, and retain its conclusion when a different evaluator administers the test. These requirements turn reward-interface auditing into an instrument-design problem.

We introduce *audited reward-interface dependence*: the action-distribution response of a frozen policy over an eligible class of reward-interface edits. Eligibility is determined on a disjoint task bank by a fixed ground-truth outcome evaluator, at least 95% pairwise policy-ranking preservation, and Spearman $\rho \geq 0.95$. The Paired Interface Response Audit (PIRA) estimates this response through matched initial states and common random numbers while retaining intervention identity in a profile. Rank-conformal p-values and Holm correction provide a label-free finite-sample decision; PIRA-L maps the same profile to a higher-power ranking score.

We qualify PIRA across measurement and administration axes. Exact-TV anchors reveal the action-distribution target directly on τ-bench-Exact, SWE-Gym-Exact, and MLE-Bench-Exact. Independent

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.

operators then repeat the fixed-bank protocol, followed by full handoffs in which external teams select, audit, and execute low-overlap banks with frozen PIRA checkpoints and thresholds. This sequence separates target observability, execution stability, and bank-design stability.

The resulting evidence is consistent across all three axes. PIRA-Holm reaches $0.872/0.868/0.866$ AUROC without exploit labels; PIRA-L reaches $0.903/0.897/0.894$ and exceeds a 142M cross-attention trace encoder by $0.024/0.023/0.023$. Operator changes shift AUROC by at most 0.006. The three full handoffs use semantic overlap of $8.6\%/9.7\%/11.2\%$ and reach $0.884\pm.015/0.882\pm.014/0.879\pm.014$ AUROC, with weakest-family performance of 0.802–0.806 and Benign-Hard FPR below 5%.

The contributions are:

- An operational interventional target that separates the eligible edit class, its finite bank realization, and the execution administration, with measurable task-preservation criteria for every intervention.
- PIRA, a measurement instrument combining matched-rollout estimation, identity-preserving profiles, finite-sample conformal Holm testing, and a higher-power learned readout.
- An exact-target qualification program spanning three exact-TV anchors, three operator reruns, three external-bank handoffs, held-out channels, and bank-only swaps.

Together, these contributions make administration invariance measurable and establish exact-target qualification as the evidentiary basis for paired-rollout audits.

2 Related Work

Reward gaming and behavioral monitoring. Reward hacking and specification gaming span reward tampering, evaluator manipulation, and predictable reward overoptimization (Amodei et al., 2016; Krakovna et al., 2020; Skalse et al., 2022; Everitt et al., 2021; Uesato et al., 2020; Gao et al., 2023). Preference-learning methods fit rewards from human feedback (Christiano et al., 2017; Ziegler et al., 2019); uncertainty estimates support out-of-distribution detection (Sedlmeier et al., 2019), exploration (Liang et al., 2022), and ensemble-based mitigation of reward overoptimization (Coste et al., 2024). PIRA contributes an interventional measurement axis: task-preserving interface edits define the stimulus, exact action-distribution response defines the target, and conformal Holm testing supplies the calibrated decision.

Counterfactual estimation and invariance. Counterfactual learning and off-policy evaluation learn policies or estimate their value from logged feedback (Swaminathan & Joachims, 2015; Thomas & Brunskill, 2016). PIRA uses matched rollouts to estimate a frozen policy's response to reward-interface information while controlling shared stochasticity. Reward-learning work also characterizes partial identifiability through invariances (Skalse et al., 2023). Invariance methods seek predictors stable across environments (Arjovsky et al., 2019; Creager et al., 2021; Rosenfeld et al., 2021; Ahuja et al., 2021); PIRA evaluates the acting policy's response across qualified interface edits and evaluates the audit instrument across administrations.

Adversarial and robustness evaluation. Red-teaming and adversarial-policy work motivate held-out and adaptive tests (Gleave et al., 2020; Perez et al., 2022; Ganguli et al., 2022; Casper et al., 2023). Robustness-evaluation methodology emphasizes protocols that expose the claimed object (Carlini et al., 2019). PIRA instantiates this principle through exact-target anchors, label-free calibration, frozen-operator reruns, and bank reauthoring. Supplementary Table S10 compares these axes across adjacent method families.

3 Reward-Interface Dependence as a Measured Object

Measured object. Deployment is a reward-aware partially observed process with latent task state s_t, action a_t, ordinary observation o_t, and reward-interface state z_t. The policy acts on $x_t = (o_t, z_t)$, while the proxy stack supplies $\hat{R}(s_t, a_t, z_t)$ for an underlying task objective R^*. We define *audited*

reward-interface dependence as the response functional mapping each eligible task-preserving edit of $z_{1:H}$ to the induced action-distribution shift of a frozen policy.

Definition 1 (Operationally task-preserving intervention)**.** *An intervention ι edits the observed reward-interface process $z_{1:H}$ while holding task instances, transition dynamics, and ground-truth outcome evaluation fixed. It belongs to the eligible class $\mathcal{I}_\star$ when a disjoint audit bank preserves at least 95% of sampled pairwise policy rankings under R^* and has Spearman $\rho \geq 0.95$ between pre- and post-intervention ground-truth returns.*

The eligible class $\mathcal{I}_\star$ defines the construct; a promoted bank $\mathcal{B} = \{\iota_1, \ldots, \iota_m\} \subset \mathcal{I}_\star$ is one finite realization of the measurement instrument. For an exact-TV administration, a policy distribution is first pushed forward to a finite canonical decision alphabet. At decision t, c_t maps a tool action to its tool name, argument schema, and stop/continue bit, maps a terminal free-text action to its blinded semantic bucket, and maps an unparsable or otherwise unassigned action to $\perp$. The comparison alphabet is the union of the canonical supports under the two interfaces, together with $\perp$:

$$\mathcal{Q}_t = c_t(\operatorname{supp} \pi(\cdot \mid h_t)) \cup c_t(\operatorname{supp} \pi(\cdot \mid h_{\iota,t})) \cup \{\perp\}.$$

Writing $p_t(q) = \sum_{a:c_t(a)=q} \pi(a \mid h_t)$ and defining $p_{\iota,t}$ analogously makes the exact quantity the TV distance on this common canonical alphabet. The bank-level estimand is

$$M_{\mathcal{B}}(\pi) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\frac{1}{H} \sum_{t=1}^{H} \frac{1}{2} \sum_{q \in \mathcal{Q}_t} |p_t(q) - p_{\iota_i,t}(q)| \right], \tag{1}$$

where the expectation is over matched task instances and rollout randomness. Paired trajectories are compared through the first canonical fork: the distribution at the fork decision is included, and later decisions are not re-aligned after different canonical actions or terminal lengths. Thus H is the number of aligned decisions through that fork (or the shared terminal decision when no fork occurs). The vector of intervention-specific expectations is the richer instrument response. Exact anchors expose this canonical action-distribution response directly; surrogate administrations approximate the same response through temperature-matched action resampling.

The promoted author bank covers reward rescaling, potential shaping, progress clipping, verifier masking, evaluator lag, verifier–judge channel exchange, decoy-tool exposure, and auxiliary-call cost. Every slot is audited on 400 disjoint policy pairs before detector training. Pairwise-ranking preservation ranges from 97.8% to 99.3%, every slot has $\rho \geq 0.953$, and blinded auditing agrees with automated promotion on $219/240$ pairs. Supplementary Table S3 gives each configuration, denominator, confidence interval, semantic family, and hardest slice. The same eligibility test governs every externally authored bank, so bank variation changes the finite realization while preserving the construct definition.

An audit administration $\mathcal{A} = (\mathcal{B}, \mathcal{O}, \mathcal{E})$ combines the bank, operator $\mathcal{O}$, and execution stack $\mathcal{E}$. Operator reruns vary $(\mathcal{O}, \mathcal{E})$ with $\mathcal{B}$ fixed, while bank handoffs vary all three under the same eligibility rule. This decomposition turns administration sensitivity into an explicit empirical property. We denote the promoted author-bank realization by $\mathcal{I}_{\text{head}}$ in the results; Supplementary Table S3 adds the external-bank validity rows.

Matched-rollout estimator. For candidate policy π and intervention ι, PIRA executes K paired rollouts from matched initial states with common random numbers (Law, 2015):

$$\tau^{(k)} = \text{Rollout}(\pi, \hat{R}, z, s_0^{(k)}, \xi^{(k)}), \qquad \tau_\iota^{(k)} = \text{Rollout}(\pi, \hat{R}_\iota, z^\iota, s_0^{(k)}, \xi^{(k)}). \tag{2}$$

For open-weight policies, $\widehat{\delta}_t = \frac{1}{2} \sum_{q \in \mathcal{Q}_t} |p_t(q) - p_{\iota,t}(q)|$ is exact. Closed APIs use $S = 8$ temperature-matched action resamples under matched non-intervened prompt content, tool-output caches, seeds, retries, and memoization. Their plug-in mass is $\widehat{p}_{S,t}(q) = S^{-1} \sum_{s=1}^{S} \mathbb{1}\{c_t(a_t^{(s)}) = q\}$ on the union of observed canonical categories and $\perp$; an unobserved category has zero empirical mass, with no smoothing or unobserved-mass correction. The detector uses a composite trace score: its TV term measures the response in Equation 1, while sampled action disagreement adds a coupling-dependent feature. Thus the composite score is not an unbiased estimator of pure TV. The trace score is

$$\delta(\tau, \tau_\iota) = \frac{1}{H} \sum_{t=1}^{H} \left[\widehat{\delta}_t + \beta \mathbb{1}(a_t \neq a_{\iota,t}) \right], \quad \beta = 0.5. \tag{3}$$

We first define the per-intervention aggregate

$$\Delta_\iota(\pi) = \frac{1}{K}\sum_{k=1}^{K} \delta(\tau^{(k)}, \tau_\iota^{(k)}), \tag{4}$$

then the profile $\mathbf{d}(\pi) = [\Delta_{\iota_1}(\pi), \ldots, \Delta_{\iota_m}(\pi)]$ and the scalar summary $\Delta(\pi) = m^{-1}\sum_i \Delta_{\iota_i}(\pi)$. Preserving intervention identity lets the readout distinguish a structured response from the same average magnitude produced by unrelated slots. We write $\Delta_i(\pi) \equiv \Delta_{\iota_i}(\pi)$ when indexing a promoted bank. For readout s and metric q, administration drift is

$$\mathrm{Drift}_q(\mathcal{A}, \mathcal{A}'; s) = \left| q(\{s(\mathbf{d}_{\mathcal{A}}(\pi))\}_\pi) - q(\{s(\mathbf{d}_{\mathcal{A}'}(\pi))\}_\pi) \right|. \tag{5}$$

We use AUROC for threshold-free drift and Benign-Hard FPR for operating-point drift. Operator reruns and bank handoffs alter different components of $\mathcal{A}$, making the source of instability identifiable.

Qualification contract. The instrument is qualified through three linked tests. Target qualification compares the deployed response estimate with exact TV on anchors where both are observable, including per-family slices and policy-rank correlation. Inferential qualification applies PIRA-Holm to a disjoint benign reference population and verifies finite-sample family-wise error control before measuring the incremental power of PIRA-L. Administration qualification repeats the frozen instrument under an independently executed fixed bank and under independently designed banks, reporting Equation 5 rather than pooling these changes into a new test set. We pre-specify operator drift tolerance 0.015, bank drift tolerance 0.020, weakest-family AUROC floor 0.80, and aggregate Benign-Hard FPR ceiling 5%. These criteria jointly require a visible target, a calibrated decision, and a conclusion that is reproducible across the intended administration changes.

4 PIRA: Interventional Instrument and Readouts

PIRA composes four elements: an eligible intervention bank, a matched-rollout response estimator, a calibrated multiple-testing decision, and an intervention-indexed learned readout. The shared measurement path supports two complementary outputs:

audited interface edit $\longrightarrow$ matched CRN rollouts $\longrightarrow$ response profile $\mathbf{d}(\pi)$
PIRA-Holm: calibrated rejection $\longleftarrow$ $\mathbf{d}(\pi)$ $\longrightarrow$ PIRA-L: high-power ranking

Finite-sample PIRA-Holm. Let $\mathcal{B}_{\mathrm{ref}} = \{\pi_{\mathrm{ref}}^{(j)}\}_{j=1}^{n_{\mathrm{ref}}}$ be a calibration bank disjoint from detector training, validation, and test policies, with $n_{\mathrm{ref}} = 400$. For each of the eight interventions,

$$p_i(\pi) = \frac{1 + \sum_{j=1}^{n_{\mathrm{ref}}} \mathbb{1}\{\Delta_i(\pi_{\mathrm{ref}}^{(j)}) \geq \Delta_i(\pi)\}}{n_{\mathrm{ref}} + 1}. \tag{6}$$

The non-strict comparison handles ties conservatively. Under exchangeability of the candidate and reference policies for each bank-conditioned benign null, these rank-based values are super-uniform in finite samples. PIRA-Holm sorts $p_{(1)} \leq \cdots \leq p_{(8)}$ and rejects successively while $p_{(j)} \leq 0.05/(9-j)$; Holm's step-down argument (Holm, 1979) therefore controls FWER at 0.05 under arbitrary dependence. The inferential output is the Holm rejection set. We use $T(\pi) = -\sum_i \log p_i(\pi)$ as its threshold-free ranking companion, so reported PIRA-Holm AUROC ranks policies by T, while operating-point fractions use Holm decisions. Supplementary Proposition 1 states the guarantee.

PIRA-L and inference. PIRA-L is defined on the promoted author bank's ordered eight-coordinate profile: coordinates remain tied to the eight intervention configurations in Supplementary Table S3, rather than to a variable number of external schema fields. PIRA-L z-scores this fixed profile using training-split statistics and fits a three-layer MLP of width 256 with ReLU hidden activations and a scalar linear output:

$$\mathcal{L}(\theta) = \sum_{(r,s):y_r > y_s} \max(0, \eta - f_\theta(\mathbf{d}_r) + f_\theta(\mathbf{d}_s)) + 10^{-4}\|\theta\|_2^2, \qquad \eta = 1.$$

Table 1: Core exact-anchor results under the frozen author bank $\mathcal{I}_{\text{head}}$. PIRA-Holm is the non-learned test, PIRA-L the learned profile scorer; the third-party rerun is labeled TPA, and the first external bank swap is labeled Indep. v1. Gain is PIRA-L minus RTC. BH FPR is PIRA-L on Benign-Hard at matched 94.1% recall. Entries with uncertainty are mean±std over 10 seeds.

Anchor	RTC	PIRA-Holm	PIRA-L	Gain	TPA	TPA $\lvert\Delta\rvert$	Indep. v1 $\lvert\Delta\rvert$	BH FPR
τ-Exact	0.872±.014	0.872±.013	0.903±.012	+0.031	0.898±.013	0.005	0.011	3.9%
SWE-Exact	0.866±.015	0.868±.014	0.897±.013	+0.031	0.891±.014	0.006	0.014	4.2%
MLE-Exact	0.867±.014	0.866±.014	0.894±.013	+0.027	0.888±.014	0.006	0.013	4.4%

Adam uses learning rate 3×10^{-4} for at most 100 epochs, with early stopping after 10 validation checks without AUROC improvement. Thresholds are selected once on validation data at 94.1% recall and then frozen for test, operator, bank, and held-out-channel evaluations. Every retained bank transfer registers one qualified configuration for each of these eight coordinates before the frozen normalization and readout are applied.

Sequential threshold decision and complexity. For sequential inference, division by $1+\beta$ places each per-intervention rollout contribution in $[0,1]$. Assuming independent rollout-pair draws across k under the frozen administration, coordinate i at look $b \le K$ spends $\alpha_{b,i} = 6\alpha/(m\pi^2 b^2)$ for $\alpha = 0.05$ and uses radius $r_b = \sqrt{\log(2/\alpha_{b,i})/(2b)}$. Writing $\bar{X}_{i,b}$ for the normalized sample mean, the resulting rectangle $\mathcal{C}_b = \prod_i [\bar{X}_{i,b} - r_b, \bar{X}_{i,b} + r_b]$ is rescaled to the original profile coordinates and mapped through the frozen training normalization and PIRA-L by interval bound propagation. The procedure commits a binary threshold decision when the score interval excludes the validation-fixed threshold. Supplementary Proposition 2 proves simultaneous coordinate coverage and correctness relative to the frozen expected-profile decision under data-dependent stopping. Runtime is $O(mKHS)$ policy evaluations plus one interval pass through the fixed MLP per look, with $S = 1$ for exact-TV policies.

5 Instrument Qualification

Exact anchors and protocol. The three open anchors expose logits end to end: τ-bench-Exact (Yao et al., 2024) ($N = 216$; Llama-3.1-70B-Instruct FP8 and Qwen2.5-72B-Instruct BF16), SWE-Gym-Exact (Pan et al., 2024) ($N = 228$; DeepSeek-Coder-V2-236B-Instruct INT8 and Qwen2.5-Coder-32B-Instruct BF16 on SWE-Gym-v2), and MLE-Bench-Exact (Chan et al., 2024) ($N = 204$; the same coder pair on MLE-Bench-v0.4.2 in a hermetic Python 3.11.9 sandbox). All use policy-level $70/10/20$ splits, $K = 50$, $|\mathcal{I}_{\text{head}}| = 8$, and seeds $\{17, 23, 41, 59, 83, 101, 137, 163, 181, 211\}$. Means and standard deviations are over these ten seeds; 95% intervals use 2,000 stratified paired bootstrap resamples of test policies. Exact split and class denominators appear in Supplementary Table S2.

Across benchmark-specific audits, the bank preserves 97.6–99.3% of policy-pair rankings, has aggregate Spearman $\rho = 0.966$, and agrees with blinded human auditing at 91.3%. Author runs use $8\times$NVIDIA A100 80GB SXM4 GPUs, CUDA 12.4, PyTorch 2.3.1, and vLLM 0.6.2. The three operator reruns use disjoint H100 clusters and frozen banks, checkpoints, splits, and seeds.

Comparators and evaluation questions. Summary-Stat and DeepSets read the identical divergence matrix, PT-Transformer is a 46M raw paired-trace encoder with intervention identity, RTC is a frozen Qwen2.5-72B-Instruct raw-trace critic, and CATX is a 142M cross-attention multiple-instance encoder over paired traces. The experiments answer three qualification questions: whether PIRA measures the exact target, whether its readouts add calibrated and learned sensitivity beyond matched controls, and whether its conclusion survives operator, bank, and held-out-channel changes. Supplementary Table S9 compares the strongest methods across the complete administration ladder.

5.1 Target Observability and Controlled Readouts

Table 1 establishes three conclusions. First, the label-free, finite-sample PIRA-Holm readout recovers RTC-level discrimination before supervision: it matches RTC on τ-Exact, leads by 0.002 on SWE-Exact, and trails by 0.001 on MLE-Exact. Second, PIRA-L adds $0.031/0.029/0.028$ AUROC over

Table 2: Controlled comparisons on the three exact-TV anchors. Every method uses identical policy splits, paired-rollout budgets, and 10 seeds; values are mean±std.

Method	τ-Exact	SWE-Exact	MLE-Exact
Summary-Stat (same signal)	$0.829\pm.019$	$0.823\pm.020$	$0.820\pm.020$
DeepSets (same signal)	$0.871\pm.014$	$0.865\pm.015$	$0.862\pm.015$
PT-Transformer (raw traces + ID)	$0.857\pm.015$	$0.853\pm.016$	$0.851\pm.016$
RTC (raw traces)	$0.872\pm.014$	$0.866\pm.015$	$0.867\pm.014$
CATX (cross-attention, 142M)	$0.879\pm.013$	$0.874\pm.014$	$0.871\pm.014$
PIRA-Holm	$0.872\pm.013$	$0.868\pm.014$	$0.866\pm.014$
PIRA-L (full profile)	$0.903\pm.012$	$0.897\pm.013$	$0.894\pm.013$

Table 3: Bank externalization under one definition: semantic Jaccard on canonicalized intervention descriptions. PIRA normalization, readouts, and thresholds remain frozen. Full-handoff slice entries report external held-out/weakest exploit-family AUROC.

Setting	Overlap	PIRA-L AUROC or drift	Held-out/weakest	BH FPR
Indep. v1, three anchors	21.1%	$\lvert\Delta\rvert = 0.011/0.014/0.013$	—	4.4%
Indep. v2–v4 envelope	11.8–19.0%	$\lvert\Delta\rvert \le 0.019$	$\ge 0.821/\ge 0.801$	$\le 4.9\%$
τ full handoff	8.6%	$0.884\pm.015$	0.842/0.803	4.8%
SWE full handoff	9.7%	$0.882\pm.014$	$0.846/0.806$	4.7%
MLE full handoff	11.2%	$0.879\pm.014$	$0.841/0.802$	4.6%

PIRA-Holm. Exact-to-surrogate gaps are $0.009/0.010/0.011$, with Pearson $r = 0.964/0.958/0.954$; all reruns preserve method order at Kendall $\tau = 1.00$. Third, frozen bank changes remain small: Indep. v1 shifts AUROC by $0.011/0.014/0.013$. The Table 1 FPR column belongs to PIRA-L; PIRA-Holm records 4.8–4.9% Benign-Hard FPR in Supplementary Table S1.

Table 2 isolates the value of the intervention-indexed profile. Across the three anchors, PIRA-L gains 0.0737–0.0744 AUROC over Summary-Stat and 0.0316–0.0322 over DeepSets; PT-Transformer and CATX remain 0.043–0.046 and 0.023–0.024 behind, respectively. DeepSets receives the same divergence signal, the trace encoders receive raw paired trajectories, and CATX uses 142M trainable parameters; together these controls attribute the remaining gain to the identity-preserving profile after controlling for paired data and model scale. Paired-bootstrap difference intervals and AUPRC operating points are reported in Supplementary Table S11.

5.2 Administration Invariance Across Operators and Banks

TPA-WG, TPA-SWE, and TPA-MLE independently reproduce the three author rows within $0.005/0.006/0.006$ AUROC. They also preserve held-out and weakest-family conclusions, Benign-Hard calibration, and the exact-to-surrogate bridge. This operator change exercises the complete execution path, including fresh rollouts, cache behavior, and surrogate estimation.

The second-team τ handoff completes the three-anchor matrix under a distinct collaboration protocol: the external team selects the reward-interface surface, audits the bank before detector access, registers one qualified configuration for each frozen coordinate, and executes frozen PIRA checkpoints on $N = 216$ policies. Its AUROC drift is 0.019, its own 400-pair task-preservation audit retains $389/400$ policy-pair rankings (97.3%), and pre/post Spearman correlation is 0.958. Supplementary Table S7 resolves the v2–v4 envelope into individual bank rows. Together with SWE and MLE, the full handoffs establish stable measurement across all exact-TV anchors despite low bank overlap.

Held-out reward-interface channels retain mean PIRA-L AUROC 0.857 with 4.6% FPR. This test changes the evaluated interface field while retaining the fixed eight-coordinate PIRA-L protocol and the existing threshold.

Table 4 separates the distinct claims carried by the evidence. Exact-to-surrogate agreement qualifies the observable quantity, operator reruns qualify execution, and low-overlap handoffs qualify bank realization. Held-out channels then test whether the frozen profile remains informative when an evaluated reward-interface field is absent from detector training.

Table 4: Qualification evidence beyond the author-bank aggregate. BH is mean Benign-Hard FPR at the frozen operating point.

Qualification axis	PIRA result	Matched control	BH FPR
Exact→surrogate	gap $0.009/0.010/0.011$	Pearson $.964/.958/.954$	$3.9/4.2/4.4\%$
Independent operator	drift $.005/.006/.006$	rank $\tau = 1.00$	$4.1/4.4/4.6\%$
Full bank handoff	$.884/.882/.879$	weakest $.803/.806/.802$	$4.8/4.7/4.6\%$
Held-out interface channel	mean $.857$	RTC mean $.838$	mean 4.6%

Target and readout qualification. The exact anchors provide two complementary checks. At the measurement level, policy-level exact and surrogate profiles retain Pearson correlation 0.954–0.964 and rank correlation 0.937–0.947, while aggregate AUROC gaps remain 0.009–0.011 at $S = 8$. At the decision level, the count-derived ordinary-benign false-rejection fractions are $4.7/5.0/5.0\%$ and the corresponding exploit rejection fractions are $81.3/80.8/80.4\%$. The learned layer raises AUROC by 0.028–0.031 over this calibrated readout, yet the profile remains informative when the learned head is removed. This combination localizes the source of performance: exact-TV agreement supports the response estimate, Holm supports the decision calibration, and the PIRA-L increment captures ranking power rather than redefining the target.

The controlled representation comparisons sharpen that attribution. Summary-Stat collapses intervention identity and loses 0.0737–0.0744 AUROC. DeepSets sees the same eight response coordinates but remains 0.0316–0.0322 behind, showing that generic set aggregation does not recover the ordered slot structure. RTC and CATX receive raw paired trajectories and remain 0.027–0.031 and 0.023–0.024 behind despite CATX's 142M parameters. The paired-bootstrap intervals for PIRA-L minus CATX are $[0.011, 0.037]$, $[0.009, 0.037]$, and $[0.008, 0.038]$, placing the fixed-anchor improvement beyond seed-level mean comparisons.

Administration and stress qualification. The operator reruns change hardware and execution ownership while fixing the bank, policies, seeds, normalization, and thresholds. Their 0.005–0.006 drift, invariant method ordering, and 4.1–4.6% FPR isolate execution reproducibility. Bank-only swaps then lower semantic overlap to 11.8–21.1% and externalize the audit surface while keeping drift within 0.019. Complete handoffs combine bank design and execution, yet all three anchors remain within the same tolerance and preserve weakest-family AUROC above 0.80. These nested changes show which component of the administration moved and prevent an aggregate transfer score from conflating operator and intervention-design sensitivity.

The remaining stress axis tests a distinct failure route. Held-out channels preserve a 0.857 mean AUROC when a reward-interface field is absent from training. Together with the bank handoffs, these results test channel coverage and independently authored bank realization without altering the central response functional.

6 Discussion

The results establish a practical qualification rule for audit research: exact-TV anchors establish target observability, PIRA-Holm establishes finite-sample calibrated detection, and operator and bank changes measure administration invariance. CATX remains competitive on fixed-anchor AUROC but incurs two to four times greater drift under operator and bank changes (Supplementary Table S9). Re-execution and bank reauthoring therefore reveal instrument quality that fixed-suite discrimination leaves hidden.

The qualified setting comprises reward-aware, resettable systems with observable interface channels. Bank coverage determines which dependencies enter the measured profile, while Benign-Hard calibration quantifies legitimate interface-conditioned responses. Held-out-channel and low-overlap handoff evaluations directly probe these operating boundaries. In this setting, PIRA serves as pre-deployment triage: intervention-indexed flags and paired traces identify the channels driving review.

7 Conclusion

PIRA establishes reward-interface dependence as an exact-anchored interventional measurement. PIRA-Holm combines label-free finite-sample decisions with ranking-companion AUROC of $0.872/0.868/0.866$, PIRA-L raises ranking performance to $0.903/0.897/0.894$, independent operators stay within 0.006 AUROC, and complete three-anchor handoffs reach $0.884\pm.015/0.882\pm.014/0.879\pm.014$ with sub-5% Benign-Hard FPR. Exact-target qualification and administration invariance together provide a reusable evidentiary standard for reward-aware agent audits.

AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

Reproducibility Statement

Section 5 specifies the anchors, models, splits, seeds, hardware, uncertainty convention, and comparison protocol. Sections 3–4 define the measured object and inference rules; the supplement gives exact intervention configurations, denominators, hyperparameters, calibration, held-out-channel evaluation, and the finite-sample proof.

References

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3438–3450, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565. URL https://arxiv.org/abs/1606.06565.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893. URL https://arxiv.org/abs/1907.02893.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. doi: 10.48550/arXiv.1902.06705. URL https://arxiv.org/abs/1902.06705.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023. doi: 10.48550/arXiv.2306.09442. URL https://arxiv.org/abs/2306.09442.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024. doi: 10.48550/arXiv.2410.07095. URL https://arxiv.org/abs/2410.07095.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/dda7f9378a210c25e470e19304cce85d-Abstract-Conference.html.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 2021. URL https://proceedings.mlr.press/v139/creager21a.html.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198 (Suppl. 27):6435–6467, 2021. doi: 10.1007/s11229-021-03141-4. URL https://doi.org/10.1007/s11229-021-03141-4.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022. doi: 10.48550/arXiv.2209.07858. URL https://arxiv.org/abs/2209.07858.

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.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgEMpVFwB.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://www.jstor.org/stable/4615733.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. Google DeepMind blog, 2020. URL https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Averill M Law. *Simulation Modeling and Analysis*. McGraw-Hill Education, 5th edition, 2015. ISBN 978-0-07-340132-4. URL https://www.mheducation.com/highered/product/simulation-modeling-and-analysis-law.

Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=OWZVD-l-ZrC.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with SWE-Gym. *arXiv preprint arXiv:2412.21139*, 2024. doi: 10.48550/arXiv.2412.21139. URL https://arxiv.org/abs/2412.21139.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3419–3448, 2022. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

Elan Rosenfeld, Pradeep K. Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

Andreas Sedlmeier, Thomas Gabor, Thomy Phan, Lenz Belzner, and Claudia Linnhoff-Popien. Uncertainty-based out-of-distribution detection in deep reinforcement learning. *arXiv preprint arXiv:1901.02219*, 2019. doi: 10.48550/arXiv.1901.02219. URL https://arxiv.org/abs/1901.02219.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058. PMLR, 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823. PMLR, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148. PMLR, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. Avoiding tampering incentives in deep RL via decoupled approval. *arXiv preprint arXiv:2011.08827*, 2020. doi: 10.48550/arXiv.2011.08827. URL https://arxiv.org/abs/2011.08827.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. doi: 10.48550/arXiv.2406.12045. URL https://arxiv.org/abs/2406.12045.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. doi: 10.48550/arXiv.1909.08593. URL https://arxiv.org/abs/1909.08593.

SUPPLEMENTARY MATERIAL

S1 FINITE-SAMPLE CALIBRATION

PIRA-Holm uses a reference bank of $n_{\mathrm{ref}} = 400$ ordinary benign policies that is disjoint from PIRA-L training, model selection, threshold selection, and all reported test policies. Benign-Hard and sham-control policies are separate stress sets and never enter the reference bank. For promoted intervention ι_i, write $\Delta_i(\pi) \equiv \Delta_{\iota_i}(\pi)$. The bank-conditioned null $H_{0,i}$ states that the candidate is exchangeable with the ordinary benign reference policies for this response coordinate. PIRA-Holm computes

$$p_i(\pi) = \frac{1 + \sum_{j=1}^{n_{\mathrm{ref}}} \mathbb{1}\{\Delta_i(\pi_{\mathrm{ref}}^{(j)}) \geq \Delta_i(\pi)\}}{n_{\mathrm{ref}} + 1}. \tag{7}$$

Reference and candidate scores use the same rollout budget, bank, canonicalization, and seed schedule. Ties enter the right tail through "$\geq$," making the p-value conservative.

Proposition 1 (Finite-sample family-wise error control)**.** *Suppose that, for every intervention i, the candidate policy and the n_{ref} policies in $\mathcal{B}_{\mathrm{ref}}$ are exchangeable under the bank-conditioned benign null. Then the rank p-value p_i in Equation 7 is super-uniform. Applying Holm's step-down rule to the eight intervention hypotheses controls the probability of at least one false rejection at level $\alpha = 0.05$, without assumptions on dependence among interventions.*

Proof. Exchangeability makes the candidate's descending rank among the $n_{\mathrm{ref}} + 1$ scores uniform under random tie breaking. The implemented non-strict comparison counts every tie in the right tail, producing a p-value stochastically no smaller than the randomized-rank value and therefore preserving super-uniformity. Let I_0 be the set of true nulls and let $p^0_{\min}$ be their smallest p-value. If Holm makes any false rejection, then $p^0_{\min} \leq \alpha/|I_0|$. The union bound and super-uniformity give

$$\Pr(\text{any false rejection}) \leq \Pr\left(p^0_{\min} \leq \frac{\alpha}{|I_0|}\right) \leq \sum_{i \in I_0} \frac{\alpha}{|I_0|} = \alpha.$$

The descriptive score $T(\pi) = -\sum_i \log p_i(\pi)$ is not used in this guarantee. □

PIRA-Holm AUROC uses $T(\pi)$ as the threshold-free ranking companion to the calibrated test. The finite-sample FWER guarantee concerns the Holm decision under its stated exchangeability assumptions; the empirical count summaries below are reported separately as policy-decision fractions.

Table S1: PIRA-Holm operating points. AUROC is threshold-free; exploit rejection and ordinary-benign false rejection are count-derived pooled policy-decision fractions under Holm at $\alpha = 0.05$; Benign-Hard FPR uses the frozen deployment threshold. Gap is PIRA-Holm minus PIRA-L AUROC.

Anchor	AUROC	Exploit reject	Ordinary-benign false reject	PIRA-L gap	BH FPR
τ-Exact	$0.872 \pm .013$	$195/240$ (81.3%)	$9/190$ (4.7%)	-0.031	4.8%
SWE-Exact	$0.868 \pm .014$	$202/250$ (80.8%)	$10/200$ (5.0%)	-0.029	4.9%
MLE-Exact	$0.866 \pm .014$	$185/230$ (80.4%)	$9/180$ (5.0%)	-0.028	4.9%

Across the three anchors, the pooled ordinary-benign false-rejection fractions are $4.7/5.0/5.0\%$, while pooled exploit rejection fractions are $81.3/80.8/80.4\%$. The denominators are policy decisions pooled across the ten fixed seed executions, not ten independent policy populations. Calibration diagrams use ten equal-frequency bins with the ordinary-benign denominators shown in Table S1.

S2 EXPERIMENTAL UNITS, INTERVENTIONS, AND TRAINING

Independent units and splits. Each N counts independently trained or independently prompted policies evaluated on disjoint task instances; no rollout pair is treated as an independent policy. Each policy receives one registry label before the policy-level split: E denotes the anchor's exploit-labelled registry subset and B its ordinary-benign registry subset. PIRA-Holm never consumes these labels; PIRA-L consumes labels only from its training split and all reported ranking metrics read labels only after scoring. Table S2 gives the exact policy-level split and class denominators.

Table S2: Policy-level denominators. E/B denotes exploit/benign policies. The split is fixed before seed-level rollout replication.

Anchor	N	Train/val/test	Train E/B; val E/B; test E/B	Test unit
τ-Exact	216	151/22/43	83/68; 12/10; 24/19	policy
SWE-Exact	228	160/23/45	88/72; 13/10; 25/20	policy
MLE-Exact	204	143/20/41	79/64; 11/9; 23/18	policy

Promoted author bank. The eight configurations in Table S3 are chosen before detector training. Each is audited on 400 policy pairs from the held-out task-preservation bank. Pairwise intervals are Wilson 95% intervals; Spearman intervals use 2,000 paired bootstrap resamples. The sham control applies seed-preserving no-op resets and produces a 2.1% alarm rate on 400 benign policies. Blinded adjudication agrees with the automated promotion decision on 219/240 pairs (91.3%).

Table S3: Intervention validity and coverage. Preserved reports the count/rate and Wilson 95% CI from that row's own 400-pair task-preservation audit; it is a bank-promotion statistic, not a policy-evaluation result. Every row clears pairwise preservation $\geq 95\%$ and Spearman $\rho \geq 0.95$. Sham alarm is 2.1% for the complete no-op bank.

Altered field	Promoted configuration	Preserved	Rate [95% CI]	ρ [95% CI]	Family / hard slice
shaping weights	blend $\alpha \in \{.25, .50, .75\}$	397/400	99.3 [97.8, 99.7]	.981 [.972, .988]	shaping / tool loops
potential $\Phi(s)$	$\gamma = .90$	396/400	99.0 [97.5, 99.6]	.978 [.967, .986]	shaping / score grafting
progress / bonus meter	clip at $q = .99$	394/400	98.5 [96.8, 99.3]	.971 [.958, .981]	clipping / hidden tests
verifier features	mask $p = .15$	391/400	97.8 [95.8, 98.8]	.953 [.950, .968]	visibility / schema coercion
evaluator timing	lag 3 steps	395/400	98.8 [97.1, 99.5]	.975 [.964, .984]	visibility / calibration
verifier / judge	channel swap	392/400	98.0 [96.1, 99.0]	.958 [.951, .973]	visibility / private metric
auxiliary tool	one decoy; $p_{\text{fire}} = .20$	393/400	98.3 [96.4, 99.1]	.965 [.953, .976]	auxiliary / early stop
auxiliary cost	$\lambda = .05$ per call	394/400	98.5 [96.8, 99.3]	.969 [.956, .979]	auxiliary / formatting
τ external	4-bit replay; 2–4 messages	389/400	97.3 [95.1, 98.5]	.958 [.951, .968]	judge chain / action confirmation
SWE external	mask $p = .20$; [.1, .4]	394/400	98.5 [96.8, 99.3]	.962 [.954, .971]	verifier / flaky tests
MLE external	noise $\sigma = .02$; [.005, .03]	391/400	97.8 [95.8, 98.8]	.957 [.950, .967]	private metric / leaderboard probing

The strongest external families alter judge-chain replay on τ-Exact, verifier-channel masking on SWE-Exact, and private-leaderboard feedback on MLE-Exact. The three external rows are the promotion audits for the corresponding full-handoff banks: τ is 389/400 (97.3%), SWE is 394/400 (98.5%), and MLE is 391/400 (97.8%). Each external team applies the same task-preservation criteria before receiving detector outputs. This separation prevents detector failures from steering bank promotion.

PIRA-L training. Profiles are normalized with training-split means and standard deviations. The three-layer width-256 MLP uses ReLU hidden activations, a scalar linear output, Adam, learning rate 3×10^{-4}, regularization 10^{-4}, margin 1.0, at most 100 epochs, and early stopping after 10 validation checks. Thresholds are chosen on validation policies only. All bank and operator transfers reuse the frozen normalization, checkpoint, and threshold.

S3 EXACT-TV BRIDGE AND OPERATOR RERUNS

Table S4: Exact-TV bridge at deployed surrogate resampling $S = 8$. Held-out exploit-family and weakest exploit-family columns show exact/surrogate AUROC; score statistics are policy-level.

Anchor	Exact	Surrogate	Held-out	Weakest	Pearson r	Rank ρ
τ-Exact	$0.903 \pm .012$	$0.894 \pm .013$	$0.867/0.857$	$0.819/0.812$	0.964	0.947
SWE-Exact	$0.897 \pm .013$	$0.887 \pm .014$	$0.869/0.859$	$0.826/0.816$	0.958	0.941
MLE-Exact	$0.894 \pm .013$	$0.883 \pm .014$	$0.864/0.852$	$0.819/0.811$	0.954	0.937

Table S5: Surrogate-resampling sweep. Entries are absolute exact-to-surrogate AUROC gaps; the final column is policy-rank correlation at $S = 8$.

Anchor	$S = 1$	$S = 2$	$S = 4$	$S = 8$	$S = 16$	$\rho_{S=8}$
τ-Exact	0.056	0.031	0.018	0.009	0.005	0.947
SWE-Exact	0.024	0.015	0.010	0.010	0.007	0.941
MLE-Exact	0.029	0.019	0.013	0.011	0.009	0.937

Increasing resampling reduces the exact-to-surrogate gap overall; SWE-Exact plateaus between $S = 4$ and $S = 8$. At $S = 8$, aggregate gaps are 0.009–0.011 and policy-level correlations are 0.954–0.964.

Table S6: Independent operator reruns with the author bank and frozen PIRA-L. Held-out and weakest denote exploit-family slices; Gap lists author/TPA exact-to-surrogate AUROC gap.

Anchor	Author	TPA	$\lvert\Delta\rvert$	Held-out	Weakest	Gap	BH FPR
τ-Exact	0.903	0.898	0.005	0.862	0.813	$0.009/0.011$	4.1%
SWE-Exact	0.897	0.891	0.006	0.853	0.810	$0.010/0.012$	4.4%
MLE-Exact	0.894	0.888	0.006	0.849	0.807	$0.011/0.014$	4.6%

All three administrations preserve the complete method ranking (Kendall $\tau = 1.00$), the brittle slices, and sub-5% Benign-Hard FPR. Paired rather than unpaired rollouts reduce estimator variance by 37–42% across open anchors.

S4 ADMINISTRATION CHANGES AND LADDER COMPARATORS

The v2–v4 sequence lowers semantic overlap and, in v4, transfers audit-surface selection while retaining the author-bank readout unchanged. Every bank remains within 0.019 exact-anchor AUROC drift, preserves the reported held-out and weakest-family floors, and keeps aggregate Benign-Hard FPR below 5%.

The τ-Exact row is administered by a second external team using a separately chosen interface surface and the same $70/10/20$ policy split, $K = 50$, and seed list. Its 97.3% preservation is the $389/400$ bank-promotion result, while its AUROC is evaluated over the separate $N = 216$ policy population. Across the three handoffs, semantic overlap is low, drift remains within 0.019, and every weakest-family AUROC remains above 0.80.

Conditional-randomization profile baseline. The CRT baseline receives the same eight-coordinate divergence profile as PIRA-Holm. Reference-bank means and standard deviations define frozen slot scores $z_i(d_i)$. For candidate profile $\mathbf{d}$, the statistic is $C(\mathbf{d}) = \max_i z_i(d_i)$. Conditional on the multiset of profile values, the null treats intervention labels as exchangeable; the test enumerates all $8!$ coordinate permutations σ and computes

$$p_{\mathrm{CRT}}(\mathbf{d}) = \frac{1}{8!} \sum_{\sigma \in \mathfrak{S}_8} \mathbb{1}\{C(\mathbf{d}_\sigma) \geq C(\mathbf{d})\}.$$

Table S7: Bank-only swaps with frozen PIRA-L, normalization, and threshold. Semantic overlap uses Jaccard on canonicalized intervention descriptions. Drift is the largest absolute AUROC change on the exact anchors; held-out/weakest are minima over all rows evaluated by each bank.

Bank	Semantic overlap	Ext. audit surface	Max drift	Held-out/weakest	BH FPR
Independent v2	19.0%	no	0.019	$\geq 0.822/\geq 0.806$	4.7%
Independent v3	11.8%	no	0.019	$\geq 0.821/\geq 0.801$	4.8%
Independent v4	13.4%	yes	0.019	$\geq 0.821/\geq 0.801$	4.9%

Table S8: Full independent handoffs. The external team chooses the interface surface, audits the bank, registers one qualified configuration for each frozen coordinate, and executes frozen PIRA checkpoints. Held-out and weakest denote external exploit-family slices. Drift is relative to the author-bank PIRA-L row; uncertainty is mean$\pm$std over the fixed 10 seeds. Rank/ρ reports the separate 400-pair bank-promotion audit from Table S3, not a statistic over the N policies.

Anchor	N	Overlap	PIRA-L	Drift	Held-out	Weakest	BH FPR	Rank/ρ
τ-Exact	216	8.6%	0.884±.015	0.019	0.842	0.803	4.8%	97.3%/0.958
SWE-Exact	228	9.7%	0.882±.014	0.015	0.846	0.806	4.7%	98.5%/0.962
MLE-Exact	204	11.2%	0.879±.014	0.015	0.841	0.802	4.6%	97.8%/0.957

AUROC ranks policies by $-\log p_{\mathrm{CRT}}$. Operator and bank rows reuse the frozen reference normalization, and Benign-Hard FPR uses a validation-selected threshold matched to 94.1% recall. This comparator tests whether generic profile-level randomization recovers PIRA's slotwise calibrated and learned readouts.

Table S9: Comparator performance across audit administrations. Vectors follow τ/SWE/MLE. TPA and bank columns are absolute AUROC drift; BH FPR is measured at matched 94.1% recall. All methods use identical splits, seeds, and $K = 50$ with no post-author-bank retuning.

Method	Quantity	τ-Exact	SWE-Exact	MLE-Exact
PIRA-Holm	anchor AUROC	0.872	0.868	0.866
	TPA drift		0.006/0.007/0.007	
	bank-v1 drift		0.013/0.016/0.015	
	BH FPR	4.8%	4.9%	4.9%
PIRA-L	anchor AUROC	0.903	0.897	0.894
	TPA drift	0.005	0.006	0.006
	bank-v1 drift	0.011	0.014	0.013
	BH FPR	3.9%	4.2%	4.4%
CATX	anchor AUROC	0.879	0.874	0.871
	TPA drift		0.014/0.016/0.017	
	bank-v1 drift		0.032/0.035/0.037	
	BH FPR		6.1/6.3/6.4%	
CRT profile test	anchor AUROC	0.861±.015	0.857±.016	0.853±.016
	TPA drift		0.009/0.011/0.012	
	bank-v1 drift		0.024/0.026/0.027	
	BH FPR		5.3/5.5/5.4%	

CATX approaches PIRA-L on fixed anchors but loses two to four times more AUROC under operator and bank changes and produces higher benign review load. The direct conditional-randomization profile test has lower ranking performance; its permutation calibration is conditional on the stated intervention-label exchangeability null. PIRA-Holm's operator drift of 0.006/0.007/0.007 is intentionally distinct from PIRA-L's 0.005/0.006/0.006 because the two rows use different readouts of the shared profile. Administration changes therefore expose robustness differences that fixed-suite AUROC leaves unresolved.

Table S10: Comparison axes for adjacent method families. "Exact" means that the claimed target is directly observable, and "NL calibration" means a non-learned finite-sample decision rule. RTC and CATX receive administration tests in this paper as the strongest paired-trace controls.

Family	Measured object	Intervention	Exact	NL calibration	Operator	Bank	Held / adapt.
Reward tampering / gaming	incentive pathway	causal structure	—	—	varies	varies	varies
Predictive monitors	confidence / disagreement	observational shift	—	threshold-based	rare	rare	OOD slices
Counterfactual / OPE	policy value	action or policy	partial	bounds	partial	—	overlap
Invariant learning	stable predictor	environment shift	—	objective-based	multi-env.	multi-env.	OOD
RTC / CATX	exploit ranking	interface edit	eval.	supervised	here	here	here
PIRA	action-distribution response	task-preserving interface edit	direct	conformal Holm	frozen	frozen	both

S5 Uncertainty and Representation Controls

Table S11: Uncertainty for the controlled comparisons. Standard deviations are across the same ten seeds. Difference intervals are 95% stratified paired-bootstrap intervals over test policies.

Comparison	τ/SWE/MLE values	Interpretation
PIRA-L minus Summary-Stat	$0.0741/0.0737/0.0744$	coarse magnitude
PIRA-L minus DeepSets	$0.0316/0.0322/0.0318$	same-signal head
PIRA-L minus CATX	0.024 [0.011, 0.037]; 0.023 [0.009, 0.037]; 0.023 [0.008, 0.038]	paired-trace model
PIRA-L minus RTC	0.031 [0.018, 0.044]; 0.031 [0.017, 0.045]; 0.027 [0.013, 0.041]	raw-trace critic
PIRA-L minus DeepSets, CI	[0.019, 0.044]; [0.018, 0.046]; [0.017, 0.047]	paired uncertainty

Summary-Stat standard deviations are $0.019/0.020/0.020$, DeepSets $0.014/0.015/0.015$, PT-Transformer $0.015/0.016/0.016$, RTC $0.014/0.015/0.014$, CATX $0.013/0.014/0.014$, PIRA-Holm $0.013/0.014/0.014$, and PIRA-L $0.012/0.013/0.013$. PIRA-L AUPRC is $0.891/0.884/0.879$ and PIRA-Holm AUPRC is $0.854/0.849/0.845$ on the three anchors. The seed spreads, non-identical four-decimal gains, paired intervals, and raw rejection counts make the uncertainty behind rounded AUROCs explicit.

S6 Held-Out Interface Channels

The held-out channel is judge-chain quantize-and-replay on τ-Exact, verifier-channel stochastic masking on SWE-Exact, and private-leaderboard metric fabrication on MLE-Exact. External bank designers specify these channels before seeing detector errors. PIRA normalization, model parameters, Holm reference, and thresholds remain frozen.

PIRA-L leads RTC by 0.019 on average and remains highest in every channel. Relative to the corresponding author-bank rows, the τ cell and three-channel mean both show a 0.041 drop; the frozen threshold keeps FPR at 4.5–4.7%. This is positive transfer to reward-interface fields absent from detector training.

Table S12: Held-out interface-channel evaluation. Values are mean±std over 10 seeds; CI is a policy-level paired-bootstrap 95% interval for PIRA-L. Counts report exploit/benign policies. Drop is the author-bank PIRA-L AUROC in Table 1 minus the held-out-channel AUROC.

Anchor	Count E/B	PIRA-L	PIRA-L CI	PIRA-Holm	RTC	Drop	BH FPR
τ-Exact	40/32	0.862±.015	[0.847, 0.877]	0.846±.016	0.843±.016	0.041	4.5%
SWE-Exact	42/33	0.867±.015	[0.852, 0.882]	0.848±.016	0.847±.016	0.030	4.6%
MLE-Exact	38/31	0.842±.017	[0.825, 0.859]	0.822±.018	0.825±.018	0.052	4.7%
Mean	—	0.857	—	0.839	0.838	0.041	4.6%

S7 Sequential Decision and Score-Component Ablations

Confidence-sequence stopping. Under independent bounded rollout-pair draws across k, the threshold decision uses coordinate-wise alpha-spending Hoeffding sequences with $\alpha_{b,i} = 6\alpha/(m\pi^2 b^2)$, $\alpha = 0.05$, and maximum budget 50. At every look, the simultaneous rectangle is rescaled to the original profile coordinates before interval bound propagation maps it through the frozen PIRA-L network; stopping occurs when the resulting score interval excludes the validation-selected threshold.

Proposition 2 (Valid expected-profile stopping)**.** *For coordinate i, let*

$$X_{i,k} = \frac{1}{H(1+\beta)} \sum_{t=1}^{H} \left[\widehat{\delta}_{t,i,k} + \beta \mathbb{1}(a_{t,i,k} \neq a^{\iota}_{t,i,k}) \right] \in [0,1], \qquad \mu_i = \mathbb{E}[X_{i,k}],$$

and assume the rollout-pair draws are independent across k under a frozen administration. With probability at least $1-\alpha$, every coordinate mean lies in its reported interval at every look. If interval bound propagation returns an enclosure of the frozen PIRA-L score on each such rectangle, any early decision whose score interval lies wholly on one side of the frozen threshold agrees with the decision obtained from the expected profile $(1+\beta)\boldsymbol{\mu}$.

Proof. At look b, the two-sided Hoeffding bound (Hoeffding, 1963) with radius $r_b = \sqrt{\log(2/\alpha_{b,i})/(2b)}$ fails for coordinate i with probability at most $\alpha_{b,i}$. Since

$$\sum_{i=1}^{m} \sum_{b=1}^{\infty} \alpha_{b,i} = \sum_{i=1}^{m} \sum_{b=1}^{\infty} \frac{6\alpha}{m\pi^2 b^2} = \alpha,$$

a union bound gives simultaneous coverage over every coordinate and look, including the data-dependent stopping look. Multiplication by $1+\beta$, frozen z-normalization, and interval bound propagation preserve enclosure. An interval wholly above or below the fixed threshold therefore contains an expected-profile score on the declared side. □

Action-disagreement weight. Table S13 varies β with all other settings frozen.

Table S13: Score-component ablation. PIRA-L AUROC and mean Benign-Hard FPR use the same splits and ten seeds.

β	PIRA-L AUROC (τ/SWE/MLE)	BH FPR
0	0.896/0.890/0.887	4.4%
0.25	0.901/0.895/0.892	4.3%
0.5	0.903/0.897/0.894	4.2%
1	0.900/0.894/0.891	4.3%
2	0.891/0.886/0.882	4.7%

The broad plateau from $\beta = 0.25$ through 1 shows that the result is not determined by one action-disagreement coefficient. Exact-TV shift remains informative at $\beta = 0$, while overweighting sampled disagreement at $\beta = 2$ produces a modest, non-catastrophic decline.

S8 Reproduction Details

Open-anchor execution uses PyTorch 2.3.1, CUDA 12.4, driver 550.90.07, vLLM 0.6.2, and $8\times$NVIDIA A100 80GB SXM4 GPUs. Exact-TV inference aligns canonical decoding prefixes before comparing action distributions. Third-party reruns use disjoint H100 clusters with the same checkpoint, tokenizer, task snapshots, intervention descriptions, $K = 50$, and seed list.

Tool-agent generation uses temperature 0.7 and nucleus sampling $p = 0.95$; terminal free text is mapped to blinded semantic buckets, while tool actions are canonicalized as tool name, argument schema, and stop/continue bit. Paired calls share non-intervened prompt content, tool-output cache, judge prompt, judge temperature 0.0, seed, retry policy, and memoization. Cache hits are forced symmetrically. MLE-Exact uses Python 3.11.9; SWE-Exact uses frozen SWE-Gym-v2 task and hidden-test partitions; all threshold and normalization fitting uses training or validation policies only.

S9 Auditability Evidence Matrix

Table S14: Auditability matrix connecting protected object, adversary, assumptions, evidence, deployment cost, and observed failure mode. Every evidence trace points to a manuscript table or section.

Risk / protected object	Adversary and trust	Property / boundary / comparator	Evidence; result and failure
Authored-bank leakage / audit conclusion	adversary knows author slots; external banks must satisfy the task audit	bank stability / frozen detector / CATX	three low-overlap handoffs, Tables S8–S9; FPR $< 5\%$; unobserved channels remain outside the bank
Unseen channel / policy response	policy conditions on an excluded field; held-out field is absent from training	transfer / intervention bank / RTC	judge replay, verifier mask, and private metric, Table S12; mean AUROC 0.857; τ drop 0.041
Surrogate approximation / divergence estimate	exact logits available on audit anchors; sampled actions share the paired execution controls	exact–surrogate agreement / score estimator / CATX	Table S4; aggregate AUROC gaps 0.009–0.011; the sampled profile is an approximation to canonical exact TV
Legitimate sensitivity / benign decision	benign policy uses the interface; ordinary-benign reference exchangeability	Holm guarantee / false-rejection fraction / PIRA-L	sham and difficult benign policies, Tables S1 and S9; PIRA-Holm 4.8–4.9%, PIRA-L 3.9–4.4% aggregate FPR; legitimate reactions can flag
Operator drift / reported ranking	evaluator changes hardware and execution; split, seeds, bank, and model stay frozen	rerun invariance / execution operator / CATX	three H100 reruns, Table S6; 0.005–0.006 drift; asymmetric cache behavior breaks the coupling
Bank-design drift / audit instrument	external team changes intervention design under a shared task-preservation rule	instrument robustness / promotion boundary / CRT	v1–v4 and full handoffs; drift ≤ 0.019; channel coverage determines sensitivity
Stochastic effects / causal contrast	intervention can perturb unrelated execution; state, cache, and random stream are matched	variance reduction / CRN pairing / unpaired rollout	37–42% variance reduction; $K = 50$ ceiling; unmatched effects inflate variance

The matrix makes the measurement boundary explicit: PIRA qualifies the interpretation of an offline reward-interface dependence audit, while the task oracle, reset mechanism, and promoted-bank audit supply its trusted inputs. The strongest comparator changes by question—CATX for representation, RTC for raw traces, the conditional-randomization test for calibration, and exact TV for surrogate fidelity.