# MONITORABILITY IS NOT PORTABILITY: WHEN CAN RUNTIME MONITORS BE REUSED?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Runtime monitors are typically validated for a particular agent configuration, yet deployment can change the model, tools, permissions, or task environment. A monitor may continue to produce confident scores after such changes even when the ranking that made those scores useful no longer transfers. The central question is therefore not whether the monitor still runs, but whether its evidence remains valid on the new target. Across GPT 4.1 mini, GPT 4.1, and GPT 5.5, HarnessAudit remains strongly monitorable, while ToolSandbox to HarnessAudit portability is 0.520, 0.822, and 0.421. Reversing the GPT 5.5 transfer gives 0.815, and a separate transfer falls significantly below chance. These results reveal a separation between monitorability and portability. The target can retain a clear harm signal even when a source trained monitor no longer preserves the same ordering between harmful and clean episodes. This separation changes what is required for reuse. Capability, marginal distribution shift, and the tested zero label adaptations are less informative than direct target probing on the transfer grid, while a small stratified probe gives the most accurate portability estimates among the tested alternatives. SELECT TRUST converts sequential target comparisons into an anytime valid reuse decision with false trust control under i.i.d. target episodes. Finally, an independently fitted and calibrated HarnessAudit breaker raises alarms before the audited harm locus on 91.7% of harmful evaluation episodes at 10.9% false cuts. Taken together, the results show that source validation is only the starting point for reuse. The target must remain monitorable, the source ranking must remain portable in the intended direction, and the resulting score must become actionable in time.

## 1 INTRODUCTION

A runtime monitor is usually validated together with the agent configuration that generated its data. Deployment may then change the base model, tool surface, permissions, memory, orchestration, or task environment. The monitor will continue to return a score, but source validation establishes only that this score was useful on the source. The deployment question is different: *does the same score still rank harmful and clean trajectories correctly on the new target?*

Recent work has shown that partial agent trajectories can support early failure prediction, online auditing, and repair across web, software, dialog, and multi-agent settings (Zhang et al., 2026; Pan et al., 2026; Baidya et al., 2026; Wang et al., 2026; Dubey, 2026; Cho et al., 2026). These results establish that useful runtime signals can exist. Reuse raises a different question: what remains valid when a monitor is carried from one deployment setting to another? Three properties matter. *Monitorability* asks whether target harm is visible from the runtime representation. *Portability* asks whether a particular source monitor preserves its ranking of harmful and clean episodes on that target. *Actionability* asks whether the resulting score reaches a usable operating point before the audited harm occurs. Conflating these properties under a single notion of robustness obscures both the source of failure and the evidence needed to respond to it.

The central empirical finding is a separation between monitorability and portability. HarnessAudit remains strongly monitorable across GPT 4.1 mini, GPT 4.1, and GPT 5.5, while ToolSandbox to HarnessAudit portability is 0.520, 0.822, and 0.421. The preregistered monotone model ordering therefore fails on this grid: the middle model transfers strongly while its two neighbours lie near chance. Transfer direction changes the result again. For GPT 5.5, reversing the same environment



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.

pair raises portability to 0.815. A separate FIN plus ToolSandbox to HarnessAudit transfer falls significantly below chance at 0.346 with 95% CI $[0.233, 0.467]$. The target can therefore retain a strong harm signal even when a ranking learned on the source becomes weak, direction dependent, or reversed.

Once monitorability and portability are separated, the next question is whether portability can be inferred without measuring the target ranking directly. The tested evidence becomes progressively more informative, but none resolves the problem as accurately as direct target measurement. Model capability does not recover the observed ordering. Marginal distribution shift cannot by itself represent the strong asymmetry between opposite transfer directions, while the tested directional coverage statistic remains too uncertain to determine target AUROC. Removing environment identity features substantially improves one difficult transfer, identifying a concrete source of shortcutting, but leaves the reuse decision unresolved. A class conditional diagnostic adds information about how harmful and clean populations move, yet a small stratified target probe gives lower portability error on the same held-out comparisons. On the transfer grid studied here, the most reliable evidence for whether a ranking survives is therefore evidence that directly measures that ranking on the target.

Direct measurement answers how portable the monitor appears to be, but deployment also requires a decision about when the evidence is sufficient. SELECT TRUST addresses this step. Fresh harmful and clean target episodes are compared pairwise, so each observation measures the same ordering that defines AUROC. The resulting evidence is accumulated through an e-process that can be inspected after every observation while retaining false trust control under i.i.d. target episodes. A zero label directional coverage screen handles clear low gap cases, while unresolved targets proceed to the labelled sequential stage. This establishes whether reuse is supported, but a portable ranking is still not an intervention policy. On a natural HarnessAudit target, an independently fitted and calibrated breaker raises alarms before the audited harm locus on 91.7% of harmful evaluation episodes at a 10.9% false cut rate. The resulting evidence chain therefore moves from whether harm is visible, to whether the source ranking survives, to whether that ranking can support timely action.

**Contributions.** The paper makes three contributions. First, it identifies and quantifies the separation between target monitorability and directional portability, showing that a strong target signal can coexist with weak or reversed source rankings. Second, it establishes a target evidence pathway for deciding reuse: direct probing is more accurate than the tested capability, shift, adaptation, and class conditional alternatives for estimating portability, and SELECT TRUST converts sequential target comparisons into an anytime valid reuse decision. Third, it separates portability from actionability and evaluates an independently calibrated breaker on natural HarnessAudit failures, completing the path from detecting a usable target signal to validating reuse and timely intervention.

## 2 Three Questions Behind Monitor Reuse

Consider an episode $\tau = (e_1, \ldots, e_T)$ with evaluator-side harm label $Y \in \{0, 1\}$. A monitor observes the runtime prefix $\tau_{1:t}$ and returns score $s_t \in \mathbb{R}$, with larger values indicating greater predicted harm. For target $T$ and source $S$, we use

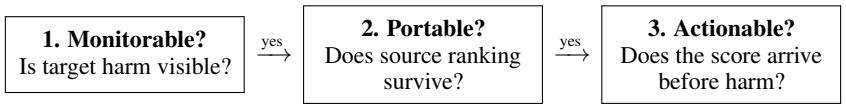
$$\begin{aligned} \mathcal{M}(T) &:= \text{target AUROC when the monitor is trained and evaluated on } T, \\ \mathcal{P}(S \to T) &:= \text{target AUROC of a monitor fitted on } S, \\ \mathcal{A}_T(\alpha) &:= \text{fraction of harmful episodes with a strict pre-locus alarm at false-cut budget } \alpha. \end{aligned}$$

These quantities form a hierarchy rather than three interchangeable metrics. $\mathcal{M}$ asks whether the target runtime view contains a learnable harm signal. $\mathcal{P}$ holds the source monitor fixed and asks whether its ranking survives one particular deployment direction. $\mathcal{A}$ adds the operating threshold and timing required for intervention. A monitor can therefore pass one stage and fail the next.

AUROC is the natural object for the first two stages because portability is an ordering question. With half credit for ties,

$$\mathrm{AUC}_D(s) = \mathbb{P}\big(s(Z_D^+) > s(Z_D^-)\big) + \tfrac{1}{2}\mathbb{P}\big(s(Z_D^+) = s(Z_D^-)\big),$$

where $Z_D^+$ and $Z_D^-$ are harmful and clean episodes from domain $D$. High target monitorability means that some monitor trained on $T$ can separate these populations. High portability is stricter: the score




| | |
|---|---|
| HarnessAudit monitorability | AUROC 0.83–0.97 across the three tested models |
| ToolSandbox → HarnessAudit | 0.520 \| 0.822 \| 0.421 |
| GPT 5.5 reverse direction | HarnessAudit → ToolSandbox: 0.815 |

Figure 1: **Monitor reuse requires three pieces of directed evidence.** The empirical contrast previews the paper's central separation: HarnessAudit remains monitorable while source-monitor portability changes sharply with model and direction.

learned on $S$ must preserve that ordering after it is carried to $T$. When portability falls below chance, the transferred score ranks the two target populations in the wrong direction.

The distinction is not only semantic. Appendix C shows constructively that source AUROC leaves target AUROC unidentified without assumptions linking the two domains. Under class-conditional stability, the gap is controlled by how the harmful and clean score distributions move separately. This is why a marginal distance can be informative yet still miss the deployment quantity: it summarizes how the pool moves before asking whether the harmful–clean ordering survives.

Direction matters for the same reason. If $\mathcal{P}(S \to T) \neq \mathcal{P}(T \to S)$, a symmetric statistic receives the same input for two different deployment outcomes. Directional statistics can recover some of that information, but they remain proxies for the quantity we ultimately need: the target ranking of the source monitor. The empirical sections therefore follow the hierarchy directly—first monitorability, then portability, then actionability.

**Experimental setting.** We hold the monitor interface fixed while varying the failure setting. The study combines the published ToolSandbox benchmark (Lu et al., 2025) with FIN and HarnessAudit study environments. FIN tests point-in-time financial evidence; ToolSandbox supplies stateful prerequisite, canonicalization, and insufficient-information tasks at varied trajectory depth; HarnessAudit tests permission and resource violations in a multi-agent tool-using framework with a hidden deterministic audit. The original GPT 5.5 study contains 1,532 episodes across the three environments, and the central cross-model comparison adds GPT 4.1 and GPT 4.1 mini on ToolSandbox and HarnessAudit. Harm labels are evaluator-side and machine-checkable. ToolSandbox uses its native milestone/minefield outcomes, while FIN and HarnessAudit use deterministic study evaluators. When a causal label is established by counterfactual repair, the implicated evidence or dependency is replaced by its repaired version and the evaluator is rerun; Appendix F gives the intervention construction.

The headline monitor is the preregistered flat SEQ model, which is used unchanged across the main transfer grid. It maps value-blind trace features to a harm-oriented score. These features include tool or event type, errors, retries, step position, and runtime verifier flags. Argument values and evaluator-only harm or locus fields are excluded from the monitor input.

For the monitorability and portability analyses, SEQ is evaluated on the completed episode at $\tau = 1.0$, producing one harm score per episode for AUROC ranking. The breaker uses the same causal feature interface, but evaluates the score repeatedly on successive prefixes so that each decision depends only on information available up to that point. Future events and the final episode length are therefore unavailable to the breaker.

Each source monitor is fitted once and then reused across the corresponding target cells. Zero-label adaptations are performed without access to target labels. When target labels are used for probe-based decisions, the decision is made from the probe alone and evaluated on a disjoint holdout, with all records from the same episode kept in the same split. Appendix A provides the complete experimental protocol together with additional early-prediction results.

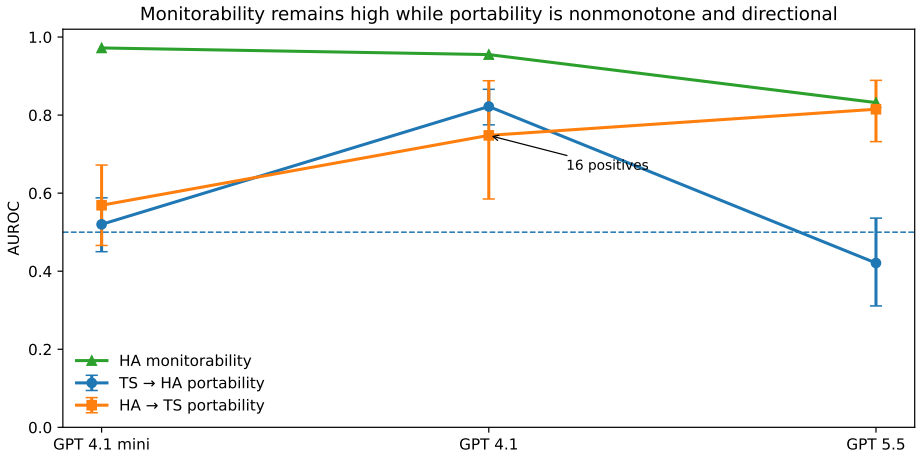



Figure 2: **The target signal survives while source-monitor portability changes.** HarnessAudit remains strongly monitorable across the three tested models, while ToolSandbox-to-HarnessAudit transfer is nonmonotone in the tested ordering. For GPT 5.5, the forward estimate is 0.421 with an episode-bootstrap 95% CI $[0.312, 0.533]$, while the reverse direction reaches 0.815.

## 3 The Target Stays Monitorable While Transfer Fails

Figure 2 shows the paper's central separation. HarnessAudit remains learnable in-domain for all three tested base models, with AUROC between 0.83 and 0.97. The target signal is therefore present in every case. Yet a monitor trained on ToolSandbox behaves very differently after transfer: HarnessAudit AUROC is 0.520, 0.822, and 0.421 for GPT 4.1 mini, GPT 4.1, and GPT 5.5. The preregistered monotone model ordering is contradicted by these three points: the middle model transfers strongly while its two neighbours lie near chance.

The effect is also directional. For GPT 5.5, ToolSandbox-to-HarnessAudit has AUROC 0.421 with episode-bootstrap 95% CI $[0.312, 0.533]$ from 31 harmful and 240 clean target episodes, whereas HarnessAudit-to-ToolSandbox reaches 0.815. The 0.421 estimate is near chance at this sample size; the evidence for an actual reversal comes from the separate composite-source cell below. What is already clear here is directionality: the base model and environment pair are fixed, but exchanging source and target changes the transferred ranking from weak to strong.

A separate composite-source experiment shows that the ranking can genuinely reverse. Training on FIN plus ToolSandbox and testing on GPT 5.5 HarnessAudit gives AUROC 0.346 with 95% CI $[0.233, 0.467]$. This cell uses a different source from the 0.421 result and becomes our diagnostic case study. Removing environment-identity tool and event features raises target AUROC to 0.589 without HarnessAudit labels. The ablation identifies one concrete failure mode: the source monitor can rely on identity cues that stop supporting the same harmful–clean ordering on the target.

The obvious shortcut is to repair the representation and stop there. That is not enough on this cell. CORAL, environment-adversarial feature removal, a structural encoder, and their combination all leave the target below the fixed useful-transfer bar (Sun and Saenko, 2016; Zhao et al., 2019). The representation ablation therefore serves as diagnosis rather than certification: it recovers part of the lost ranking, but the monitor that will actually be deployed still has to be checked on the target. The central phenomenon is now sharp: HarnessAudit remains monitorable, while the usefulness of a source score depends on how that score travels across the source–target boundary.

This separation also tells us what evidence to collect next. A weak $\mathcal{M}(T)$ points to the runtime view or monitor class. Strong $\mathcal{M}(T)$ with weak $\mathcal{P}(S \to T)$ points to the source–target transfer. Strong portability then leaves a separate operating question about $\mathcal{A}_T(\alpha)$. The experiments below ask how much target evidence is needed to resolve the middle stage and how that evidence can be turned into a controlled reuse decision.

Table 1: **Direct target probing gives the most accurate portability estimate in the frozen comparison.** Probe error uses a complement reference: target $J = 2\,\mathrm{AUC} - 1$ is recomputed from episodes outside each probe draw.

| Evidence | Observed result | What the result says |
|---|---|---|
| Model capability | TS→HA: 0.520, 0.822, 0.421 | Capability ordering does not recover portability on this grid |
| Marginal shift | RBF MMD $\rho = -0.35$; linear MMD $\rho = -0.22$ | Marginal distance has a weak, uncertain relation to target ranking |
| Directional coverage | $\rho = -0.38$ overall; stronger point estimate on the original representation | Direction helps, but the proxy remains insufficient for the target AUROC |
| Zero-label adaptation | Identity-feature removal moves 0.346 to 0.589 | Representation repair recovers part of one failure |
| Diagnostic vs. probe | MAE 0.297/0.458 vs. 0.184/0.189 | Direct target ranking gives lower held-out error |

# 4 Target Evidence Beats Transfer Shortcuts

## 4.1 Can portability be inferred without direct measurement?

The separation in Section 3 raises a practical question: can the target ranking be predicted without measuring it directly? We test an increasingly informed sequence of shortcuts on the same transfer grid. Model capability uses no target data. Marginal shift looks at unlabeled source and target feature pools. Zero-label adaptation changes the representation. Class-conditional diagnostics add target labels but still infer portability through intermediate summaries. The direct probe is the final step: it measures the harmful–clean ordering that defines target AUROC itself.

The marginal-shift results make the directional issue concrete. The preregistered RBF MMD relation to transfer has Spearman $\rho = -0.35$ with 95% CI $[-0.68, 0.06]$, and linear MMD gives $-0.22$ with CI $[-0.54, 0.15]$. The post-hoc directional coverage gap has the strongest point estimate, $-0.38$ overall and $-0.56$ on the original representation, while its intervals still span zero. These statistics are useful diagnostics of how source and target differ, but their uncertainty is large relative to the decision we need to make about one source ranking on one target.

The reason is visible in the AUROC decomposition. The bound from Section 2 depends on separate harmful and clean shifts. A single marginal discrepancy averages over those classes, while portability asks whether their pairwise ordering survives. The class-conditional diagnostic therefore adds target identifiability, class-conditional signature shift, monitor polarity, and balanced marginal shift. It is a substantially richer predictor than MMD alone, and it uses the same target labels that define the class structure.

Even with that extra information, direct measurement is more accurate on the frozen held-out comparisons. On the common scale $J = 2\,\mathrm{AUC} - 1$, the diagnostic ridge has MAE 0.297 under model holdout and 0.458 under family holdout. A single stratified probe draws up to eight harmful and twelve clean target episodes. Recomputing the reference $J$ on the unselected complement removes overlap between the probe sample and its evaluation reference; the resulting probe MAE is 0.184 and 0.189. The diagnostic error is therefore 1.6 and 2.4 times larger on the same scale. The result is conceptually simple: the probe observes the very pairwise ordering that defines portability, whereas the diagnostics infer that ordering through intermediate summaries.

The diagnostic becomes more useful once target evidence is available. Monitor polarity agrees with the sign of $J$ in 93% of cells, and flipping diagnosed reversed monitors raises holdout $J$ by a median 0.699 across five reversed cells. The diagnostic is therefore better suited to deciding how to react to a measured failure than to replacing the measurement itself. Appendix D gives the full decomposition and complement-reference analysis; Appendix J reports a sealed four-direction DeepSeek V3.2 check.

## 4.2 Turning target evidence into a reuse decision

A target probe gives an estimate; deployment needs a rule for deciding when that evidence is sufficient. SELECT TRUST tests whether target AUROC exceeds the fixed useful-portability bar $b = 0.70$. The

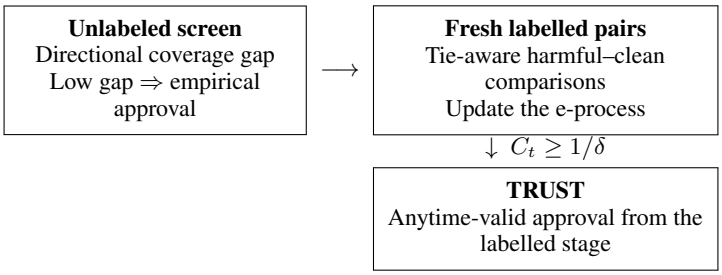



Figure 3: **SELECT TRUST separates free screening from controlled sequential approval.** The unlabeled screen expands coverage for low-gap targets; unresolved targets enter the labelled stage, whose e-process provides the false-trust guarantee.

procedure has two stages. First, an unlabeled directional-coverage screen computes

$$g(S,T) = 1 - \frac{1}{|T|}\sum_{z\in T}\max_{x\in S} k(z,x),$$

and uses a leave-one-cell-out calibrated low-gap threshold to approve easy cases at zero-label cost. Other targets proceed to the labelled stage. The screen is one-sided by design: low gap can cheaply identify an easy reuse case, while larger gaps remain candidates for direct measurement because several high-gap transfers in the grid still retain useful ranking.

For the labelled stage, each step forms one fresh harmful–clean pair and records the tie-aware comparison

$$X_t = \mathbb{1}\{s(Z_t^+) > s(Z_t^-)\} + \tfrac{1}{2}\mathbb{1}\{s(Z_t^+) = s(Z_t^-)\}.$$

Under $H_0 : \mathrm{AUC}_T(s) \leq b$, SELECT TRUST combines these comparisons with a predictable aGRAPA bet to form

$$C_t = \prod_{i=1}^{t}\left(1 + \lambda_i(X_i - b)\right), \qquad C_0 = 1,$$

with $\lambda_i$ determined only by earlier pairs. Reuse is approved when $C_t \geq 1/\delta$. Intuitively, each pair asks the same question as AUROC: did the source monitor score the harmful target episode above the clean one? When the true target ranking is no better than $b$, those pairwise wins cannot sustain uncontrolled capital growth; when the ranking is stronger, evidence can accumulate until the approval boundary is crossed.

**Theorem 1 (Anytime false-trust control)** *Assume target episodes are i.i.d. draws from the deployment population, each sequential comparison uses one fresh harmful episode and one fresh clean episode, and $\lambda_t \in [0, 0.9/b]$ is predictable. If the target satisfies $\mathrm{AUC}_T(s) \leq b$, then the probability that the labelled SELECT TRUST stage ever approves reuse is at most $\delta$.*

The proof is short: under the null, $C_t$ is a nonnegative supermartingale, and Ville's inequality gives $\mathbb{P}(\sup_t C_t \geq 1/\delta) \leq \delta$. Appendix E gives the canonical aGRAPA rule used for the reported table and the full argument. The practical consequence is optional stopping. The audit may inspect the evidence after every new pair and stop as soon as the trust bar is crossed, while the false-trust statement remains tied to the same $\delta$.

We evaluate SELECT TRUST on 17 transfer cells, ten above $b$ and seven below it, with $\delta = 0.10$. At pair cap 10, the labelled stage alone reaches sensitivity 0.200 with empirical false trust 0.002. Adding the screen raises sensitivity to 0.551, reduces mean use to about 13.6 distinct labelled episodes, and records empirical false trust 0.000. Pair caps and label budgets are different quantities: each comparison consumes one harmful and one clean episode, while cells approved by the screen consume no labels. Label acquisition is therefore governed by the rarer target class as well as by the sequential stopping time.

The matched-budget comparison clarifies the purpose of the sequential rule. Fixed-sample Hanley–McNeil approves more above-threshold cells, and the raw point rule is more sensitive still. The sequential procedure instead concentrates the gain on the safety axis: false trust is 0.000 at roughly 14

Table 2: **At matched label budgets, SELECT TRUST trades approval coverage for false-trust control.** Sensitivity is approval among the ten above-threshold cells; false trust uses all 17 cells as its denominator.

| Rule | Mean labels | Sensitivity | False trust |
|---|---|---|---|
| Screen + sequential, cap 10 | 13.6 | 0.551 | **0.000** |
| Screen + fixed H–M, $k = 18$ | 12.7 | 0.642 | 0.014 |
| Screen + fixed H–M, $k = 20$ | 14.1 | 0.645 | 0.008 |
| Screen + point, $k = 14$ | 9.9 | **0.917** | 0.071 |

Table 3: **Independent calibration closes the final stage of the evidence chain.** The scorer and threshold are fitted on the calibration half; all rates below are measured on the disjoint evaluation half.

| Target | Harmful eval episodes | Strict pre-locus alarms | False cuts |
|---|---|---|---|
| HarnessAudit natural | 24 | **0.917** | 0.109 |
| ToolSandbox study | 22 | 0.318 | **0.086** |

labels and the approval rule remains valid under optional stopping. The fixed-sample Hanley–McNeil baseline reaches false trust 0.005 only around 21 labels in the matched comparison, while the point rule has no corresponding sequential guarantee. Reusing episodes across all harmful–clean pairs creates a seemingly efficient alternative, but that dependence raises empirical false trust to 0.118. Fresh sequential evidence is therefore part of the operating design, not a bookkeeping detail.

## 5 Portability Is Not Protection

A portable ranking is still not a guardrail. Deployment needs a threshold, and that threshold has to be crossed before the audited harm locus rather than after it. AUROC cannot answer either question: two monitors can rank complete episodes equally well while concentrating their score increases at very different points in the trajectory. Actionability therefore has to be calibrated on the target after portability has been established.

HarnessAudit provides the cleanest end-to-end setting. Its natural arm contains permission and resource violations produced under the default permissive configuration, without injected faults. The GPT 4.1 mini target contains 591 episodes with 38 harmful outcomes and is strongly monitorable in-domain at AUROC 0.887. Source monitors also rank this pool well: a configured GPT 4.1 mini monitor reaches 0.949 and a GPT 5.5 study monitor reaches 0.932 under simultaneous model and configuration shift. This lets us hold the ranking question largely fixed and ask the operational one: can a causal-prefix score cross a usable threshold early enough?

The target is also operationally interesting because process harm can be hidden by ordinary completion metrics. In the HarnessAudit rule-only analysis, 74 of 79 harmful episodes preserve nominal utility. A successful-looking task outcome can therefore coexist with a permission or resource violation. This is precisely the setting in which an early process monitor adds information that terminal utility alone cannot supply.

For the main breaker evaluation, episodes are split into calibration and evaluation halves. The SEQ scorer is fitted using the calibration half, its threshold is selected there, and the disjoint evaluation half is scored without refitting. A strict alarm must occur before the audited harm locus; any alarm on a clean episode is a false cut. On HarnessAudit, the breaker raises strict pre-locus alarms on 91.7% of the 24 harmful evaluation episodes, with a 10.9% false-cut rate over 275 clean episodes. On ToolSandbox, the corresponding strict-alarm rate is 31.8% over 22 harmful episodes at 8.6% false cuts over 35 clean episodes. The contrast shows why actionability cannot be inherited from portability: even after ranking quality has been established, alarm timing remains a property of the target process and operating threshold.

The pooled out-of-fold replay provides the full operating curve behind the point estimate. At its nominal 10% operating point, strict early alarms cover 86.8% of harmful HarnessAudit episodes and

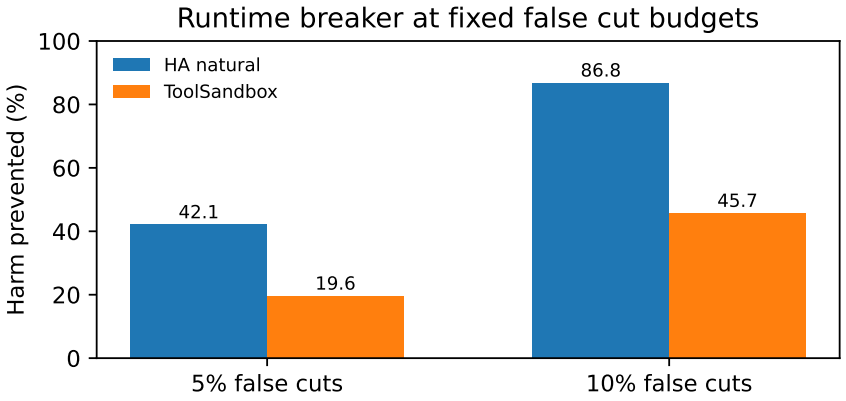



Figure 4: **The broader breaker operating curves preserve the same target ordering.** The pooled out-of-fold replay varies the false-cut operating point and records alarms relative to the audited harm locus. The independently fitted/calibrated headline result is reported in Table 3.

Table 4: **A directed evidence record connects validation to one deployment direction.**

| Evidence | What to report | Decision supported |
|---|---|---|
| Monitorability | Target in-domain performance under the runtime view | Whether the target exposes a usable harm signal |
| Portability | Source, target, direction, monitor version, target audit, target AUROC | Whether this source ranking survives on this target |
| Actionability | Alarm timing, false-cut operating point, threshold calibration, evaluation setting | Whether the ranking supports intervention before the audited harm locus |

37.0% on ToolSandbox, consistent with the independent evaluation's strong/weak target ordering. The result completes the logic of the paper: portability decides whether the source ranking is worth carrying onto the target; actionability decides whether a target-calibrated score can support an intervention before the audited violation occurs.

## 6 What Should Travel With a Monitor?

The answer is not a single source AUROC. A reused monitor should travel with a *directed evidence record*: the source checkpoint, deployment target, runtime representation, harm definition, portability audit, reuse threshold, and intervention policy under which the monitor was evaluated. The record makes direction explicit and ties the score to the conditions under which reuse was actually established.

The three stages also diagnose where reuse fails. Weak monitorability points to the observability layer: the runtime representation or monitor needs to change before transfer is meaningful. Strong monitorability with weak portability points to the source–target relationship: adaptation can be attempted, but the adapted monitor still needs target evidence before approval. Strong portability with weak actionability points to the operating layer: thresholding, score timing, or intervention policy needs calibration. These cases are operationally different even when all three would otherwise be described as "distribution shift."

For benchmark designers, the results favor reporting target monitorability alongside transfer and evaluating both directions when the deployment setting permits it. That pairing tells the reader whether a low transferred AUROC reflects a weak target signal or a source ranking that failed to travel. For monitor builders, the representation ablation shows how the record supports diagnosis: removing environment-identity features moves one difficult cell from 0.346 to 0.589, identifying a

repairable part of the shift while leaving the final reuse decision to the target audit. For deployers, SELECT TRUST governs approval and the breaker calibration sets the operating point.

The same logic determines when evidence should be refreshed. A material change in base model, tool surface, permissions, task family, monitor representation, or intervention policy creates a new deployment direction or changes what the score means. This changes the unit of versioning: not the monitor alone, but the monitor together with the conditions under which reuse was established. Runtime assurance then becomes a sequence of answerable questions with matching evidence, instead of one robustness number carrying several meanings at once.

## 7 Related Work

Runtime-monitoring work increasingly predicts failure from partial trajectories, audits agent behavior online, and triggers repair or restart (Zhang et al., 2026; Pan et al., 2026; Baidya et al., 2026; Wang et al., 2026; Dubey, 2026; Cho et al., 2026). That literature establishes the first premise needed here: agent trajectories can contain useful early-warning information. Our object is different. We start after a monitor has already been trained and validated and ask whether that particular score remains valid after the deployment condition changes. Monitorability, portability, and actionability separate target signal, transferred ranking, and intervention timing rather than treating source prediction quality as a proxy for all three.

Domain adaptation and transferability estimation provide the closest statistical analogues. Domain-adaptation theory relates source and target error through discrepancy and labeling assumptions, and invariant marginal representations can still fail when label structure changes (Ben-David et al., 2010; Zhao et al., 2019). MMD measures distribution discrepancy (Gretton et al., 2012); LEEP, LogME, and GBC estimate how well a representation or source predictor transfers to a new task (Nguyen et al., 2020; You et al., 2021; Pándy et al., 2022). Our deployment quantity is narrower and directional: the harmful–clean ranking induced by one runtime monitor on one target. This is why the experiments pair unlabeled shift summaries and adaptation with direct measurements of target AUROC rather than treating discrepancy itself as the endpoint.

SELECT TRUST connects this portability question to sequential inference. Ville's inequality and modern e-process methods support evidence that remains valid under optional stopping (Ville, 1939; Cai et al., 2026). We use that machinery for a concrete deployment error: false trust, approving reuse when target AUROC is at or below the useful-portability bar. The method therefore complements monitor architectures and adaptation methods rather than competing with them; it governs when target evidence is strong enough to reuse the resulting score. Appendix G gives a broader comparison, including formal monitor coverage under shift (Zhang, 2026).

## 8 Conclusion

Runtime monitor reuse is a target evidence problem. The central finding is a separation between monitorability and portability: a deployment target can retain a strong runtime harm signal even when a monitor learned on the source no longer preserves the same ranking on the target. Across the transfer grid, portability changes substantially with both model and transfer direction, including one verified setting in which the ranking falls below chance. The tested capability, distribution shift, and adaptation signals do not recover the target ordering as accurately as direct target evidence. A small target probe therefore provides a more faithful estimate of portability, while SELECT TRUST turns sequential target comparisons into an anytime valid reuse decision with false trust control under i.i.d. target episodes. The independently calibrated breaker completes the evidence chain by showing that portability must also translate into a score that becomes actionable before harm.

The broader consequence is a different view of monitor validation for deployment. A monitor should travel with evidence about the conditions under which its reuse has been established, rather than with a single source performance number. That evidence asks whether the target remains monitorable, whether the source ranking remains portable in the intended direction, and whether the resulting score can support timely intervention. Monitor reuse should therefore be established on the target, not inherited from the source.

## AI USE STATEMENT

Generative AI was used to improve the presentation of the paper and to assist with the implementation of experiments. The authors reviewed the AI-assisted work and verified the reported experimental implementations. Ground-truth labels in the reported experiments were produced by the stated deterministic evaluators or counterfactual procedures.

## REPRODUCIBILITY STATEMENT

All primary empirical results are produced from fixed evaluation sets with deterministic labels. For the main study, the tasks, metrics, decision margins, and analysis setup were specified before the final experiments. Additional model evaluations, representation ablations, selective-trust analyses, natural-harm analyses, and the transfer-diagnosis study are reported as follow-up analyses. We provide necessary information in the supplementary document. If the paper is accepted, we will release the experimental code, configurations, and results on github.

## REFERENCES

Avinash Baidya, Xinran Liang, Ruocheng Guo, Xiang Gao, and Kamalika Das. When evidence is sparse: Weakly supervised early failure alerting in dialogs and LLM agent trajectories. arXiv preprint arXiv:2606.05414, 2026.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Leheng Cai, Qirui Hu, and Weijia Li. Asymptotic anytime valid inference for U-statistics. arXiv preprint arXiv:2605.14692, 2026.

Seonglae Cho, Franklin Cardenoso Fernandez, Umar Mohammed, Zekun Wu, Kleyton Da Costa, Ilham Wicaksono, and Adriano Koshiyama. Automata from Agent Traces: Failure and Next Step Prediction. arXiv preprint arXiv:2608.23670, 2026.

Sunny Dubey. Real time detection and repair of LLM agent failures. arXiv preprint arXiv:2608.02464, 2026.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7294–7305, 2020.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using Bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9172–9182, 2022.

Sitong Pan, Yipeng Shen, Yilin Lu, Caiwen Ding, Lu Cheng, and Qianwen Wang. Monitoring web agents without internal signals: Observable trajectories and key step supervision. arXiv preprint arXiv:2609.02057, 2026.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, pp. 443–450, 2016.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier Villars, Paris, 1939.

Chenyu Wang, Yunbo Lyu, Junda He, Zhou Yang, Chenxing Zhong, Yaniv Harel, and David Lo. Fail fast, restart smart: Early failure prediction and restart for SWE agentic tasks. arXiv preprint arXiv:2608.03222, 2026.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pretrained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12133–12143, 2021.

Ruiyang Zhang. Why formal monitors fail: Attack distribution entropy as a coverage bound for LTL based LLM agent safety. arXiv preprint arXiv:2608.01388, 2026.

Boxuan Zhang, Jianing Zhu, Zeru Shi, Dongfang Liu, and Ruixiang Tang. AgentForesight: Online auditing for early failure prediction in multi agent systems. arXiv preprint arXiv:2605.08715, 2026.

Han Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7523–7532, 2019.

# A Additional Experimental Details

## A.1 Configuration axes and evaluation discipline

The agent configuration varies deployment-relevant components including skill or orchestration logic, memory, tool surface, verifier granularity, context management, and provenance surfacing. The FIN design uses a fractional factorial subset. A later audit shows that configuration main effects are modest relative to injected fault effects, so the main text uses "configuration transfer" for reuse across the agent stack.

The initial confirmatory study froze task pools, metrics, margins, and analysis before its final run. New model arms, representation ablations, selective-trust procedures, and external-validity tests are follow-up analyses.

## A.2 Monitor and selective-trust implementation

The headline predictor is the flat SEQ monitor fixed in the transfer study. Its value-blind trace representation records tool/event type, errors, retries, step position, and runtime verifier flags, while excluding argument values and evaluator-side harm/locus fields. Headline transfer uses the completed-episode score ($\tau = 1.0$); early-warning and breaker analyses re-evaluate the same causal feature interface on successive prefixes.

The unlabeled SELECT TRUST screen uses the asymmetric directional coverage gap

$$g(S,T) = 1 - \frac{1}{|T|}\sum_{z\in T}\max_{x\in S} k(z,x),$$

where $k$ is the kernel similarity used by the screen. The low-gap threshold is calibrated leave-one-cell-out on the remaining transfer cells and can approve a target at zero-label cost; otherwise the target proceeds to the labelled sequential stage. Screen decisions and sequential decisions are reported separately in the operating analysis.

## A.3 Early prediction and structural lead

The monitor can be evaluated at several prefix fractions. On the merged original pool, SEQ recall at a 10% false alarm rate is 0.23, 0.41, 0.61, and 0.63 using the first 25%, 50%, 75%, and 100% of the trace. The full trace AUROC is 0.898. Environment-stratified full-trace AUROC is 0.900 for FIN, 0.946 for ToolSandbox, and 0.904 for HarnessAudit.

Fault-locus design determines the available lead window. In FIN, all 406 structural breaks occur at step 0 or 1, while ToolSandbox and HarnessAudit use varied loci. Mean structural lead relative to the first behavioral alarm is +20.7 steps in ToolSandbox and +4.18 in HarnessAudit, compared with $-1.68$ in FIN; the FIN value reflects the near-immediate injection design.

# B Additional Results

## B.1 Original prediction and generalization results

Table 5: Merged original pool prediction results at the full trace. Recall is measured at a 10% false alarm rate.

| Method | Recall at 10% FA | AUROC |
|---|---|---|
| Outcome so far | 0.018 | 0.676 |
| Confidence | 0.000 | 0.500 |
| Verifier | 0.000 | 0.514 |
| GBDT | **0.750** | **0.932** |
| SEQ | 0.625 | 0.898 |
| REL | 0.393 | 0.778 |
| MON WM | 0.607 | 0.840 |
| ATTR GRAPH | 0.429 | 0.776 |

Table 6: SEQ generalization over the original frozen splits. A later ToolSandbox axis top up increases support to 37 positives and yields AUROC 0.896 with CI $[0.810, 0.966]$.

| Split | AUROC | Transfer gap | Positive support |
|---|---|---|---|
| Standard | 0.898 | 0.000 | 56 |
| Held-out configuration | 0.775 | 0.123 | 154 |
| Held-out interaction | 0.637 | 0.262 | 243 |
| Blind fault | 0.840 | 0.058 | 41 |
| Held-out time | 0.632 | 0.267 | 211 |
| Held-out company | 0.356 | 0.542 | 54 |
| Held-out family | 0.155 | 0.743 | 54 |
| Cross environment, pooled | 0.330 | 0.568 | 77 |

The configuration distance frontier for SEQ is AUROC 0.898, 0.806, and 0.767 at distance 0, 1, and 2. No collected cell lies at distance at least 3. In leave-one-environment-out evaluation, training FIN plus HarnessAudit and testing ToolSandbox gives 0.633 with CI $[0.535, 0.738]$; training FIN plus ToolSandbox and testing HarnessAudit gives 0.346 with CI $[0.233, 0.467]$. Pairwise FIN to ToolSandbox is stronger at 0.832 with CI $[0.751, 0.904]$.

## B.2 Judge audit and natural arm

On FIN, the original transcript judge has AUROC 0.153. A controlled 21-episode audit fixes the judge prompt and varies only available evidence: capped tool log, full tool log, and full log plus abstention, verifier, and outcome proxy signals produce AUROC 0.046, 0.102, and 0.060 respectively. The absolute values differ from the original run because the audit uses a single finance-specific prompt to isolate the effect of transcript detail, but every condition remains far below chance.

In the HarnessAudit rule-only silent-fracture audit, $n = 503$ with 79 harmful and 424 clean episodes. Seventy-four of 79 harmful episodes preserve nominal utility, mean utility is 0.979 for harmful and 0.952 for clean episodes, and the two one-sided tests (TOST) procedure declares equivalence at $\epsilon = 0.15$. There are 178 fracture-without-harm episodes. On the natural permissive arm, fracture rate is 235 of 591 and causal harm rate is 38 of 591; in-domain AUROC is 0.887 with CI $[0.80, 0.96]$, configured to natural transfer is 0.949 with CI $[0.90, 0.99]$, and cross-model plus configuration transfer from GPT 5.5 is 0.932 with CI $[0.88, 0.97]$.

### B.3 CHEAP MEASUREMENT BASELINES

The preregistered RBF MMD relation to transfer has Spearman $\rho = -0.35$ with CI $[-0.68, 0.06]$; linear MMD gives $-0.22$ with CI $[-0.54, 0.15]$. The post hoc asymmetric coverage gap gives $-0.38$ over all cells and $-0.56$ on the original representation, with confidence intervals still spanning zero. A simple labeled probe gives trust decision accuracy and false trust of 0.83/0.07 at 10 labels, 0.90/0.04 at 20, 0.95/0.03 at 30, and 0.98/0.02 at 50.

The final SELECT TRUST comparison uses the canonical aGRAPA betting rule on the same 17 cells. At pair cap 10, sequential-only sensitivity/false trust is 0.200/0.002; adding the unlabeled screen gives 0.551/0.000 at about 13.6 distinct labels. Matched-label fixed-sample baselines are more sensitive but less conservative: screened Hanley–McNeil gives 0.642/0.014 at 12.7 labels ($k = 18$) and 0.645/0.008 at 14.1 labels ($k = 20$), while the screened point rule gives 0.917/0.071 at 9.9 labels ($k = 14$). The sequential advantage is therefore on the false-trust axis, not raw sensitivity.

## C AUC SHIFT BOUND

**Source AUROC leaves target AUROC unidentified.** Fix the score function $s(z) = z$. For any $a, b \in [0, 1]$, let the source clean score be $0$ almost surely and the source harmful score be $1$ with probability $a$ and $-1$ otherwise. Then $\mathrm{AUC}_S(s) = a$. Independently, let the target clean score be $0$ almost surely and the target harmful score be $1$ with probability $b$ and $-1$ otherwise. Then $\mathrm{AUC}_T(s) = b$. Thus the same scoring rule can have any chosen source AUROC and any chosen target AUROC unless extra assumptions link the two domains.

For a harmful–clean score pair define

$$g_s(z^+, z^-) = \mathbb{1}\{s(z^+) > s(z^-)\} + \tfrac{1}{2}\mathbb{1}\{s(z^+) = s(z^-)\}.$$

Let $Q_D = P_D^+ \otimes P_D^-$, so $\mathrm{AUC}_D(s) = \mathbb{E}_{Q_D} g_s$. Because $0 \le g_s \le 1$,

$$|\mathbb{E}_{Q_T} g_s - \mathbb{E}_{Q_S} g_s| \le \mathrm{TV}(Q_T, Q_S).$$

For the product bound, insert $P_S^+ \otimes P_T^-$ and use the triangle inequality:

$$\begin{aligned}\mathrm{TV}(P_T^+ \otimes P_T^-, P_S^+ \otimes P_S^-) \le\ & \mathrm{TV}(P_T^+ \otimes P_T^-, P_S^+ \otimes P_T^-) \\ & + \mathrm{TV}(P_S^+ \otimes P_T^-, P_S^+ \otimes P_S^-).\end{aligned}$$

Tensoring both measures with the same probability distribution preserves total variation, yielding

$$\mathrm{TV}(P_T^+, P_S^+) + \mathrm{TV}(P_T^-, P_S^-).$$

If $P_T^+ = P_T^-$, then $Z^+$ and $Z^-$ are exchangeable. The two strict ordering events have equal probability, and tie aware AUROC is therefore $1/2$.

## D DIAGNOSTIC DECOMPOSITION AND PROBE COMPARISON

The diagnostic analysis computes feature distances after source-only standardization with an RBF kernel whose bandwidth is selected from source pairwise distances. Let $\mu_D^+$ and $\mu_D^-$ denote kernel mean embeddings of harmful and clean features in domain $D$. The four diagnostics are

$$\begin{aligned} D_T &= \|\mu_T^+ - \mu_T^-\|, \\ C_{ST} &= \tfrac{1}{2}\left(\|\mu_S^+ - \mu_T^+\| + \|\mu_S^- - \mu_T^-\|\right), \\ B_{ST} &= \Delta_S \Delta_T, \qquad \Delta_D = \mathbb{E}[s(Z) \mid Y = 1, D] - \mathbb{E}[s(Z) \mid Y = 0, D], \\ M_{ST} &= \left\| \tfrac{1}{2}(\mu_S^+ + \mu_S^-) - \tfrac{1}{2}(\mu_T^+ + \mu_T^-) \right\|. \end{aligned}$$

They represent target class separability, class-conditional signature shift, monitor polarity, and balanced marginal shift.

The frozen environment-level analysis uses 54 transfer cells. A ridge on $(D_T, C_{ST}, B_{ST})$ has MAE 0.2971 under leave-one-model-out and 0.4579 under leave-one-family-out on $J = 2\,\mathrm{AUC} - 1$. Each single probe draw samples $\min(8, n_+)$ harmful and $\min(12, n_-)$ clean episodes. Against

the original full-pool reference, mean per-draw errors are 0.162 and 0.169. Recomputing $J$ on the unselected complement gives 0.184 and 0.189; every draw leaves both classes in the complement. The corresponding diagnostic-to-probe ratios are 1.6 and 2.4, preserving the ordering under the complement reference.

The smaller 0.0551 and 0.0658 errors come from a different estimator: average 100 probe draws and then fit a one-variable ridge under the same grouped holdouts. Across 100 draws the union of sampled episodes covers essentially the full target pool, and the ridge fit also uses other training cells' full-pool $J$ values. These numbers are useful as a repeated-measurement baseline but do not represent a 20-distinct-label audit.

Polarity serves as a labelled action cue. Its sign matches $J$ in 93% of cells, and flipping diagnosed reversed monitors increases holdout $J$ by median 0.699 across five reversed cells. Computing $B_{ST}$ uses target harm and clean labels, placing it within the target-evidence stage.

| Estimator / reference | Model holdout | Family holdout |
|---|---|---|
| Diagnostic ridge | 0.2971 | 0.4579 |
| Single probe, full-pool reference | 0.162 | 0.169 |
| Single probe, complement reference | 0.184 | 0.189 |
| Ridge on average of 100 probes | 0.0551 | 0.0658 |

Figure 5: MAE on $J = 2\,\mathrm{AUC} - 1$. The complement-reference row removes the sampled probe episodes from the evaluation reference; the repeated-probe ridge uses substantially more than one 20-label draw.

## E Proof of Anytime False Trust Control

Assume target episodes are i.i.d. draws from a superpopulation and each run uses fresh, disjoint harmful–clean pairs. With half credit for ties, $X_t \in [0, 1]$ and

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mathrm{AUC}_T(s) \le b$$

under $H_0$. The implemented canonical aGRAPA bet is

$$\lambda_t = \mathrm{clip}\left( \frac{\widehat{\mu}_{t-1} - b}{\widehat{\sigma}^2_{t-1}}, 0, \frac{0.9}{b} \right),$$

where $\widehat{\mu}_{t-1}$ and $\widehat{\sigma}^2_{t-1}$ use only $X_1, \ldots, X_{t-1}$. The bet is therefore predictable and clipped to $[0, 0.9/b]$. Hence each factor $1 + \lambda_t(X_t - b)$ is nonnegative and

$$\begin{aligned} \mathbb{E}[C_t \mid \mathcal{F}_{t-1}] &= C_{t-1}\mathbb{E}[1 + \lambda_t(X_t - b) \mid \mathcal{F}_{t-1}] \\ &\le C_{t-1}. \end{aligned}$$

Thus $C_t$ is a nonnegative supermartingale with $C_0 = 1$. Ville's inequality gives

$$\mathbb{P}\left( \sup_t C_t \ge 1/\delta \right) \le \delta.$$

Approving only at such a crossing controls the probability of ever trusting under $H_0$ by $\delta$, even with adaptive stopping. The guarantee is formulated for i.i.d. deployment episodes; a fixed finite-pool formulation uses a different null correction.

## F Counterfactual Repair

The ToolSandbox repair study first screens 30 broken prerequisite scenarios, freezes the nine scenarios with baseline harm at least 0.5, and runs four paired intervention conditions. Mean harm is 0.781 at baseline, 0.750 after early repair of the responsible dependency, 0.848 after late repair of the responsible dependency, and 0.882 after an irrelevant early repair. The primary paired contrast between responsible and irrelevant repair is 0.125 with bootstrap 95% CI $[0.03, 0.25]$; the conservative two-sided permutation test gives $p = 0.25$. The ordering provides supporting evidence for the intervention analysis, while the main actionability result comes from the independently calibrated breaker.

## G Related Work in More Detail

**Runtime agent monitoring.** Recent work forecasts failures from partial trajectories in web agents, software agents, dialogs, and multi-agent systems (Zhang et al., 2026; Pan et al., 2026; Baidya et al., 2026; Wang et al., 2026). Dubey reports transfer of failure ranking across external corpora while also finding that a deployment-specific healthy null must be recalibrated (Dubey, 2026). Cho et al. recover finite-state structure from agent traces and report a gap between within-model and cross-model failure prediction (Cho et al., 2026). These studies show that transfer can occur. Our question is target-specific: whether a particular source monitor should be trusted after a model, configuration, direction, or failure family change, and how to make that decision with controlled target evidence.

**Shift, transferability, and anytime evidence.** Domain adaptation theory separates source error from distribution discrepancy, and invariant marginal representations can fail when label structure changes (Ben-David et al., 2010; Zhao et al., 2019). MMD is a standard discrepancy measure (Gretton et al., 2012); LEEP, LogME, and GBC estimate transferability from target predictions or representations (Nguyen et al., 2020; You et al., 2021; Pándy et al., 2022). Zhang studies coverage of fixed formal agent monitors under shift (Zhang, 2026). SELECT TRUST instead treats a learned monitor's target AUROC as the deployment quantity and uses an e-process with Ville's inequality (Ville, 1939). Recent work develops broader anytime-valid inference for U-statistics (Cai et al., 2026). SELECT TRUST uses these classical sequential tools for a monitor-reuse decision problem and evaluates that formulation on the transfer grid.

## H Scope Pilots on Terminal Correctness

Two gated pilots test whether the same value-blind prefix representation can forecast terminal correctness failures. A numeric constraint family has 36 episodes and in-domain AUROC 0.26. A real coding agent scored by hidden unit tests has 20 pilot episodes and AUROC 0.17. In both settings, a harmful terminal value can remain invisible until completion, so the prefix representation carries no useful process signature. Collection stops at the planned gate. We use these pilots only to delimit scope: the present monitor is suited to process harm that changes the observed trajectory, not arbitrary terminal correctness.

## I Additional Ablations and Design Diagnostics

**Flat sequence monitoring matches the relational variant.** A flat sequence monitor matches or exceeds the relational model on the main prediction tasks, and explicit configuration conditioning adds little. Headline analyses therefore use the simpler sequence representation.

**Information-flow mechanism frequency.** A held-out information-flow violation mechanism fires zero times across 1,815 single-agent HarnessAudit episodes spanning GPT 4.1 mini, GPT 4.1, and GPT 5.5. A targeted multi-agent indirect-injection probe on GPT 4.1 mini also yields zero delivered injections in 12 attempts. This mechanism therefore supplies no positive episodes for a prospective ranking test in the frozen design.

**FIN interaction resolution.** The FIN configuration study is a resolution III fractional factorial, where every two-factor interaction aliases with a main effect. The frozen corpus therefore resolves main effects rather than non-additive composition. An offline scan finds no defensible super-additive interaction after accounting for the aliasing; higher-resolution designs are the natural next test of configuration composition.

## J Prospective DeepSeek Seal

The prospective model is DeepSeek V3.2. Predictions are committed before unblinding under repository commit `7ddf346`. ToolSandbox episodes are degenerate under malformed multi-turn function calls, so the prospective check uses HarnessAudit: 67 usable episodes, with 18 harmful and 49 clean. Four frozen monitors commit to a positive direction before labels are opened. After

unblinding, all four target $J$ values are positive, between 0.22 and 0.47. The check therefore validates four positive-direction commitments on the new model.